Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)
Survey Track

Pretrained Language Models for Text Generation: A Survey

Junyi Li[1,3†], **Tianyi Tang**[2†], **Wayne Xin Zhao**[1,3*] and **Ji-Rong Wen**[1,2,3]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]School of Information, Renmin University of China
[3]Beijing Key Laboratory of Big Data Management and Analysis Methods
{lijunyi,steven_tang,jrwen}@ruc.edu.cn, batmanfly@gmail.com

Abstract

Text generation has become one of the most important yet challenging tasks in natural language processing (NLP). The resurgence of deep learning has greatly advanced this field by neural generation models, especially the paradigm of pretrained language models (PLMs). In this paper, we present an overview of the major advances achieved in the topic of PLMs for text generation. As the preliminaries, we present the general task definition and briefly describe the mainstream architectures of PLMs for text generation. As the core content, we discuss how to adapt existing PLMs to model different input data and satisfy special properties in the generated text. We further summarize several important fine-tuning strategies for text generation. Finally, we present several future directions and conclude this paper. Our survey aims to provide text generation researchers a synthesis and pointer to related research.

1 Introduction

Text generation, which is often formally referred as natural language generation, has become one of the most important yet challenging tasks in natural language processing (NLP). It aims to produce plausible and readable text in human language from input data (*e.g.,* a sequence and keywords). Researchers have developed numerous techniques for a wide range of applications of text generation [Li *et al.*, 2021a]. For example, machine translation generates text in a different language based on the source text [Yang *et al.*, 2020a]; summarization generates an abridged version of the source text to include salient information [Guan *et al.*, 2020].

With the recent resurgence of deep learning, various works have been proposed to solve text generation tasks based on recurrent neural networks (RNN) [Li *et al.*, 2019], convolutional neural networks (CNN) [Gehring *et al.*, 2017], graph neural networks (GNN) [Li *et al.*, 2020], and attention mechanism [Bahdanau *et al.*, 2015]. One of the advantages of these neural models is that they enable end-to-end learning of semantic mappings from input to output in text generation. Besides, neural models are able to learn low-dimensional, dense vectors to implicitly represent linguistic features of text, which is also useful to alleviate data sparsity.

Despite the success of neural models for text generation, a major performance bottleneck lies in the availability of large-scale datasets. Existing datasets for most of supervised text generation tasks are rather small (except machine translation). Deep neural networks usually have a large number of parameters to learn, which are likely to overfit on these small datasets and do not generalize well in practice.

In recent years, the paradigm of pretrained language models (PLMs) is thriving [Peters *et al.*, 2018]. The idea is to first pretrain the models in large-scale corpus and then finetune these models in various downstream tasks to achieve state-of-the-art results. It is widely recognized that PLMs can encode a large amount of linguistic knowledge from corpus and induce universal representations of language. Therefore, PLMs are generally beneficial for downstream tasks and can avoid training a new model from scratch [Brown *et al.*, 2020]. Moreover, with the increasing of computational power and the emergence of Transformer architecture [Vaswani *et al.*, 2017], PLMs have advanced from shallow to deep and achieved outstanding performance in many tasks, such as BERT [Devlin *et al.*, 2019] and GPT [Radford *et al.*, 2019]. Therefore, researchers have proposed various methods to solve text generation tasks based on PLMs. Pretrained on large-scale corpus, PLMs are able to understand natural language accurately and express in human language fluently, both of which are critical abilities to fulfill the text generation tasks. Existing surveys in this area have only partially reviewed some related topics. Zaib *et al.* [2020] and Guan *et al.* [2020] provided a synthesis to the research on some text generation subtasks, *i.e.,* dialogue systems and summarization, but did not go broader to the other important generation tasks. Qiu *et al.* [2020] summarized two generations of PLMs for the whole NLP domain and introduced various extensions and adaption approaches of PLMs. To the best of our knowledge, our survey is the first work that presents a comprehensive review of PLMs for text generation. It aims to provide text generation researchers a synthesis and pointer to related research.

To start with, we first present a general task definition with the formulations of different text generation tasks in Sec-

[†]Equal contribution.
[*]Corresponding author.

tion 2, and then briefly describe the mainstream architectures of PLMs that are used in text generation in Section 3. Since the core of text generation is to model the semantic mappings from input to output, we further organize the major advances with respect to the two aspects of *input* and *output* in Section 4-5. For input, we mainly discuss how to adapt existing PLMs to different data types. For output, we study how to satisfy special properties for the generated text. Furthermore, we summarize several important fine-tuning strategies for text generation in Section 6. Finally, we present several future directions and conclude this paper in Section 7.

2 Task and Typical Applications

In what follows, we formally define the text generation task. The core of text generation is to generate a sequence of discrete tokens $\mathcal{Y} = \langle y_1, \ldots, y_j, \ldots, y_n \rangle$, where each y_j is drawn from a word vocabulary $\mathcal{V}$. In most cases, text generation is conditioned on input data, such as attributes, text and structured data, which is denoted as $\mathcal{X}$. Formally, the text generation task can be described as:

$$P(\mathcal{Y}|\mathcal{X}) = P(y_1, \ldots, y_j, \ldots, y_n|\mathcal{X}). \quad (1)$$

According to input $\mathcal{X}$, we next introduce several typical applications of text generation:

• If $\mathcal{X}$ is not provided or a random noise vector $\boldsymbol{z}$, this task will degenerate into language modeling or unconditional generation task [Radford *et al.*, 2019], which aims to generate text without any constraint.

• If $\mathcal{X}$ is a set of discrete attributes (*e.g.,* topic words, sentiment labels), the task becomes topic-to-text generation or attribute-based generation [Keskar *et al.*, 2019]. The information in $\mathcal{X}$ plays the role of guiding the text generation process and controlling the modes of the generated text.

• If $\mathcal{X}$ is structured data like knowledge graph or table, this task will be considered as KG-to-text or table-to-text generation, called data-to-text generation [Li *et al.*, 2021c]. This task aims to generate descriptive text about structured data.

• If $\mathcal{X}$ is multimedia input such as image and speech, the task becomes image caption [Xia *et al.*, 2020] or speech recognition [Fan *et al.*, 2019]. The core of image caption is to generate a description of an image, while speech recognition enables programs to process human speech into a text format.

• The most common form of $\mathcal{X}$ is also a text sequence, and there exist several applications such as machine translation, summarization and dialogue system. Machine translation [Conneau and Lample, 2019] aims to translate text from one language into another language automatically, summarization [Zhang *et al.*, 2019b] is focused on generating condensed summary of a long document, and dialogue system [Wolf *et al.*, 2019] is designed to converse with humans using natural language.

We present the formulations for the major text generations in Table 1.

Input $\mathcal{X}$	Tasks
Random noise	Unconditional text generation
Discrete attributes	Topic-to-text generation
	Attribute-based generation
Structured data	Data-to-text generation
Multimedia	Image caption
	Speech recognition
Text sequence	Machine translation
	Summarization
	Dialogue system

Table 1: Major tasks and inputs for text generation.

3 Standard Architectures for Text Generation

Pretrained language models (PLMs) are pretrained with a mass of unlabelled text data and can be fine-tuned on downstream generation tasks. Pretrained on large-scale corpus, PLMs encode massive linguistic and world knowledge into vast amounts of parameters, which can enhance the understanding of language and improve the generation quality. The idea of pretraining is inspired by human beings, *i.e.,* we transfer and reuse our old knowledge of what we have learned in the past to understand new knowledge and handle a variety of new tasks. In this way, PLMs can successfully perform on new tasks with their old experience and knowledge.

Owing to the great achievements that Transformer [Vaswani *et al.*, 2017] has made, almost all PLMs employ the backbone of Transformer. For the text generation tasks, some of PLMs utilize the standard Transformer architecture following basic encoder-decoder framework, while the others apply a decoder-only Transformer. Next, we will introduce these two methods successively.

Encoder-decoder Transformer. A standard Transformer utilizes the encoder-decoder architecture, which is composed of two stacks of Transformer blocks. The encoder is fed with an input sequence, while the decoder aims to generate the output sequence based on encoder-decoder self-attention mechanism. Based on aforementioned architecture, models such as MASS [Song *et al.*, 2019], T5 [Raffel *et al.*, 2020], and BART [Lewis *et al.*, 2020] have improved quality of the generated text.

Decoder-only Transformer. Models such as GPT [Radford *et al.*, 2019; Brown *et al.*, 2020] and CTRL [Keskar *et al.*, 2019] employ a single Transformer decoder blocks, which is typically used for language modeling. They apply unidirectional self-attention masking that each token can only attend to previous tokens. Besides language modeling, several works also utilize the decoder-only achitecture to generate text conditioned on input text. However, these models do not have an independent module to encode input sequence. Interestingly, they concatenate the input and output sequence with a special token (*e.g.,* "`[SEP]`") and employ a novel seq2seq masking [Dong *et al.*, 2019] that each token in the input sentence can attend to each other and generated tokens can attend to all input tokens and previous generate ones. Compared to unidirectional masking, seq2seq masking is a natural way for decoder-only PLMs to solve conditional gen-

eration tasks, which is similar to the encoder-decoder architecture. Raffel *et al.* [2020] has researched the performance between the above two methods and made a conclusion that the addition of an explicit encoder-decoder attention is beneficial.

The core of text generation tasks is to learn the semantic mappings from input to output. On one hand, different tasks will correspond to a variety of input data, and we need to develop special techniques to model different data types. On the other hand, the generated text should satisfy important properties in order to cope with different task requirements. Next, we discuss the recent advances with respect to the two aspects, *i.e., input* and *output*.

4 Modeling Different Data Types from Input

As discussed in Section 2, different text generation tasks usually involve specific kinds of input. In this section, we will introduce three main kinds of input for text generation, *i.e.,* unstructured input, structured input, and multimedia input, and discuss how to model these input data in PLMs.

4.1 Unstructured Input

In NLP research, most of studies focus on modeling unstructured text input (*e.g.,* sentence, paragraph, and document). To generate satisfactory output text, it requires an excellent capacity of language understanding beyond surface meaning of individual words in the input text. Thus, Liu and Lapata [2019] and Zheng and Lapata [2019] employed PLMs (*e.g.,* BERT [Devlin *et al.*, 2019]) as text encoder for condensing text into low-dimensional vectors while preserving most of its meaning. Compared with traditional shallow neural models (*e.g.,* CNN), PLMs have a large number of parameters encoding massive world knowledge, which is potentially beneficial to capture the core meaning of text.

In some cases, the input text might be a long document consisting of several sentences and paragraphs. For PLMs trained on sentences or short paragraphs, they are less capable of accurately modeling long-range dependencies in a document. Considering this challenge, Zhang *et al.* [2019b] and Xu *et al.* [2020b] proposed hierarchical BERT to learn interactions between sentences with self-attention for document encoding. Besides, for capturing inter-sentential relations, DiscoBERT [Xu *et al.*, 2020a] stacked graph convolutional network (GCN) on top of BERT to model structural discourse graphs. By directly operating on the discourse units, DiscoBERT retains capacities to include more concepts or contexts, leading to more concise and informative output text.

We observe that most recent PLMs are pretrained on English text. While, many multilingual generation tasks such as machine translation involve multiple languages and certain languages are low-resource. This challenge hinders the wide application of monolingual PLMs to multilingual text generation tasks. Therefore, Conneau and Lample [2019] proposed to learn cross-lingual language models (XLMs) for multilingual language understanding. Based on cross-lingual PLMs, text generation models can still obtain effective input word embeddings even in a low-resource language [Wada and Iwata, 2018].

4.2 Structured Input

Structured data (*e.g.,* graph and table) is also a critical kind of input for text generation in many real-world applications such as weather report generation. However, in real-world scenario, it is difficult to collect a large amount of labelled structured data with ground-truth text for training. Since pretrained on large-scale corpus, PLMs encode a large amount of linguistic knowledge and show excellent few-shot capabilities in many tasks. Motivated by this, Chen *et al.* [2020b] and Gong *et al.* [2020] explored incorporating PLMs for data-to-text generation, especially in few-shot settings.

When applying PLMs to structured data, a major challenge is how to feed structured data into PLMs, which are originally designed for sequential text. To adapt to the sequential nature of PLMs, Ribeiro *et al.* [2020] and Mager *et al.* [2020] linearized input knowledge graph (KG) and abstract meaning representation (AMR) graph into a sequence of triples, Li *et al.* [2021b] introduced an additional graph encoder to encode the input KG, and Gong *et al.* [2020] employed a template-based method to serialize input table into text sequence. For example, the attribute-value pair "*name: jack reynolds*" will be serialized as a sentence "*name is jack reynolds*". However, direct linearization will lose the structural information of original data, which may lead to generating unfaithful text about data. Thus, in addition to generating faithful text, Gong *et al.* [2020] proposed an auxiliary reconstruction task for recovering the structural information of input data, which can enhance the capacity of modeling structural information.

In general, the output text should retain as much as important information from structured data. Therefore, to generate high-fidelity text adhereing to input, the pointer generator mechanism [See *et al.*, 2017] is adopted to copy words from input knowledge data [Chen *et al.*, 2020b]. Through grounding PLMs on external knowledge, it is likely to endow a generative model with both rich knowledge and good generalization ability. Besides, Gong *et al.* [2020] proposed a content matching loss for measuring the distance between the information in input data and the output text.

4.3 Multimedia Input

In addition to the above textual data, several attempts have been made to take as input multimedia data (*e.g.,* image, video, and speech) such as image caption and speech recognition. Both VideoBERT [Sun *et al.*, 2019b] and CBT [Sun *et al.*, 2019a] conducted pretraining for the video caption task. While, they performed pretraining only for the BERT-based encoder to learn bidirectional joint distributions over sequences of visual and linguistic tokens. So they have to train a separate video-to-text decoder, which tends to cause a *pretrain-finetune discrepancy*. In contrast, Unified VLP [Zhou *et al.*, 2020] used a shared multi-layer Transformer network for both encoding and decoding. Following UniLM [Dong *et al.*, 2019], they pretrained the model on two masked language modeling (MLM) tasks, like cloze tasks designed for sequence-to-sequence LM. Inspired by generative pretraining objectives in GPT, Xia *et al.* [2020] proposed a cross-modal pretrained model (XGPT) by taking images as inputs and using the image caption task as the basic generative task in the pretraining stage.

Besides image and video, speech recognition is also hungry for human-transcripted supervised data. So a number of unsupervised and semi-supervised methods are developed to integrate PLMs for weakly-supervised learning. For example, Fan *et al.* [2019] proposed an unsupervised approach to pretraining encoder-decoder model with unpaired speech and transcripts. Two pretraining stages are used to extract acoustic and linguistic information with speech and transcripts, which is useful for downstream speech recognition task.

5 Satisfying Special Properties for Output Text

In different text generation tasks, the generated text should satisfy several key properties. In this section, we will introduce three key properties in text generation, *i.e.,* relevance, faithfulness, and order-preservation.

Relevance. According to the linguistic literatures [Li *et al.*, 2021c], in text generation, *relevance* refers that the topics in output text is highly related to the input text. A representative example is the task of dialogue systems, which requires the generated response to be relevant to the input dialogue history. In addition to the dialogue history, a condition corresponding to the type of response might be also provided as an external input such as the topic of response and the persona of speaker. The generated responses should also be relevant to the condition. Recently, due to the absence of long-term memory, RNN-based models still tend to generate irrelevant output text and lack consistency with input. Therefore, through applying PLMs to the task of dialogue systems, TransferTransfo [Wolf *et al.*, 2019] and DialoGPT [Zhang *et al.*, 2020] were able to generate more relevant and context-consistent responses than traditional RNN-based models. Furthermore, to generalize to various types of conditions, Zeng and Nie [2020] utilized elaborated condition blocks to incorporate external conditions. They used BERT for both encoder and decoder by utilizing different input representations and self-attention masks to distinguish the source and target sides of dialogue. On the target (generation) side, a new attention routing mechanism is adopted to generate context-related words. Similar approaches have been used in non-conditioned dialogue [Bao *et al.*, 2020].

Faithfulness. Similarly, faithfulness is also a critical property of text, which means the content in generated text should not contradict the facts in input text. Sometimes, it further means the generated text is in accord with the world facts. A representative example is the task of text summarization, which aims to generate faithful text representing the most important information within the original content. Pretrained on large collections of text, PLMs are potentially beneficial to generate faithful text by utilizing background knowledge. Rothe *et al.* [2020] experimented with a large number of settings to initialize the encoder and decoder with three outstanding PLMs, *i.e.,* BERT, GPT and RoBERTa. With pretraining, the models are more aware of the domain characteristics and less prone to language model vulnerabilities. Consequently, they are more confident in predicting tokens from the document, hence, improving faithfulness. To improve the level of faithfulness of summary, Kryscinski *et al.* [2018] proposed to decompose the decoder into a contextual network that retrieves relevant parts of the source document and a PLM that incorporates prior knowledge about language generation. Also, to generate faithful text in different target domains, Yang *et al.* [2020b] fine-tuned PLMs on target domains through theme modeling loss. The role of the theme modeling module is to make the generated summary semantically close to the original article.

Order-preservation. In NLP area, order-preservation denotes that the order of semantic units (word, phrase, etc.) in both input and output text is consistent. The most representative example is the task of machine translation. When translating from source language to target language, keeping the order of phrases consistent in source language and target language will ensure the accuracy of the translation results to some extent. One line of research to achieve the order-preservation property is to perform semantic alignment in machine translation. Yang *et al.* [2020a] proposed Code-Switching Pre-training (CSP) for machine translation. They extracted the word-pair alignment information from the source and target language, and then applied the extracted alignment information to enhance order-preserving. Besides, it is more common to perform translation across multiple languages, called multilingual machine translation [Conneau and Lample, 2019]. However, little work can effectively enhance order-preservation for any pairs of languages. Thus, Lin *et al.* [2020] proposed mRASP, an approach to pretraining a universal multilingual machine translation model. The key to mRASP is the technique of randomly aligned substitution, which enforces words and phrases with similar meanings across multiple languages to be aligned in the representation space. Also, Wada and Iwata [2018] focused on aligning word representations of each language, making it possible to preserve the word order consistent cross multiple languages.

6 Fine-tuning Strategies for Text Generation

For text generation with PLMs, a key factor is how to design suitable fine-tuning strategies. In this part, we review several commonly-used fine-tuning strategies from different views.

6.1 Data View

When applying PLMs to text generation tasks especially in a new domain, how to design suitable and effective fine-tuning strategies adapting to the characteristics of new domain is an important consideration.

Few-shot Learning. In many text generations, it is difficult and expensive to obtain sufficient annotated data. Owing to the success of pretraining, PLMs can encode massive linguistic and world knowledge, which provides an effective solution to data scarcity. A commonly adopted approach is to plug the existing module with pretrained parameters. Then we fine-tune it with a few, one, or even no examples for the studied task, which are so-called few-shot, one-shot and zero-shot, respectively. For example in multilingual translation, some low-resource languages lack sufficient parallel corpus. XLM [Conneau and Lample, 2019] proposed to learn cross-lingual language models and can leverage the knowledge learned

Data	Categories	Methods
Input	Unstructured	BERT acts as text encoders [Liu and Lapata, 2019; Zheng and Lapata, 2019], hierarchical PLMs for document modeling [Zhang *et al.*, 2019b; Xu *et al.*, 2020b], and cross-lingual PLMs for multilingual input text [Conneau and Lample, 2019; Wada and Iwata, 2018].
	Structured	Linearize KG and AMR graph as triple sequence [Mager *et al.*, 2020; Ribeiro *et al.*, 2020], graph encoder for encoding KG [Li *et al.*, 2021b], and serialize table into template-based text sequence [Gong *et al.*, 2020].
	Multimedia	Video caption [Sun *et al.*, 2019b; Sun *et al.*, 2019a], image caption [Xia *et al.*, 2020], and speech recognition [Fan *et al.*, 2019].
Output	Relevance	Fine-tune PLMs in dialogue systems for generating more relevant and context related responses [Wolf *et al.*, 2019; Zhang *et al.*, 2020], and generalize to any type of input conditions based on BERT [Zeng and Nie, 2020].
	Faithfulness	Improve faithfulness with several PLMs [Rothe *et al.*, 2020], retrieve relevant parts from input and incorporate prior knowledge of PLMs [Kryscinski *et al.*, 2018], and generate faithful text in different target domains through theme modeling loss [Yang *et al.*, 2020b].
	Order-preservation	Word-pair alignment [Yang *et al.*, 2020a], a universal multilingual machine translation model [Lin *et al.*, 2020], and word representation alignment [Wada and Iwata, 2018].

Table 2: Categories of input types and output properties for text generation.

in high-resource languages to low-resource languages. Using the method proposed in Section 4, few-shot learning can also be applied in data-to-text tasks, such as table-to-text generation[Chen *et al.*, 2020b; Gong *et al.*, 2020] and KG-to-text generation[Li *et al.*, 2021b]. Chen *et al.* [2020b] directly fed GPT-2 with a small amount of serialized attribute-value pairs and Gong *et al.* [2020] further applied multiple tasks to better leverage structured information of tables. Moreover, Li *et al.* [2021b] proposed representation alignment to bridge the semantic gap between KG encodings and PLMs in order to enhance the correspondence between KG and text.

Domain Transfer. Equipped with vast amounts of parameters and pretrained on large-scale corpus, PLMs have powerful generalization capability. However, they still cannot directly adapt to new domains with large distribution discrepency from pretraining domain [Hendrycks *et al.*, 2020]. An effective solution is to continue training PLMs on specific data with pretraining objectives before fine-tuning them on target tasks. Mask prediction is a widely used method, attempting to predict the masked tokens using the remaining ones. There exist several variants of masking ways in domain transfer. Zeng and Nie [2020] proposed TF-IDF based masking to select more condition-related tokens to mask, in order to focus on domain features. Document masking is usually utilized in summarization task to capture document-level features of long documents [Zhang *et al.*, 2019b].

6.2 Task View

Besides characteristics of new domains, it is also meaningful to consider some special concerns such as language coherence and text fidelity in specific generation tasks when fine-tuning PLMs.

Enhancing Coherence. To enhance the language coherence, an important approach is to better model language context during fine-tuning. Models fine-tuned by contrastive learning are good at distinguishing whether a sentence pair is similar or not. Through this method, PLMs are forced to understand the positional or semantic relationship between two sentences, so that they can derive better representations. Next sentence prediction (NSP) is a commonly adopted way to judge whether two input sentences are consecutive segments, which can be applied to summarization [Yang *et al.*, 2020b] and dialog system [Wolf *et al.*, 2019]. Zheng and Lapata [2019] proposed to rearrange the sentence order according to their semantic similarities. CBT [Sun *et al.*, 2019a] proposed noise contrastive estimation (NCE) in cross-modal training to encourage the model to identify the correct video-text pair compared to a set of negative distractors. Denoising autoencoding (DAE) takes the corrupted text as input and aims to recover the original text. The model fine-tuned with DAE has a strong ability to understand the overall sentences and capture longer-range correlations. For example, TED [Yang *et al.*, 2020b] utilized DAE to refine essential semantic information for abstractive summarization. XGPT [Xia *et al.*, 2020] attempted to model the underlying text-image alignments using image-conditioned denoising autoencoding (IDA), in order to force the model to reconstruct the whole sentence.

Preserving Fidelity. Text fidelity refers that how the generated text adheres to the original input information, which is an important aspect to consider in many text generation tasks. The universal structure in PLMs is unable to retain the text fidelity in specific text generation tasks. For the table-to-text generation task, the structure information of table is required to be encoded. Gong *et al.* [2020] proposed to utilize multi-task learning, in order to reconstruct from table embeddings and enforce the match between table embeddings and content embeddings. Besides, the pointer generator [See *et al.*, 2017] can be applied to KG-to-text generation to copy the entity and relation information in KG [Chen *et al.*, 2020b].

6.3 Model View

To enhance the quality of generated text, a key is to well train the parameters of PLMs according to task-specific data, so that PLMs can capture the semantic characteristics specially for the generation task. However, as mentioned above, task-specific data is inadequate, thus it is likely to occur the overfitting case when fine-tuned on limited data. In this part, we will introduce several fine-tuning methods in view of models. Gu *et al.* [2020] employed a fixed teacher GPT to preserve the knowledge encoded in another fine-tuned GPT. Chen *et al.* [2020a] proposed to utilize a BERT model (teacher) as supervision to guide the Seq2Seq model (student) for better generation performance. Besides, Liu and Lapata [2019] utilized two optimizers to update the parameters of PLM and initial module separately, in order to solve the discrepancy between two modules. There also exist other ways to guide the fine-tuning process. For example, Reinforcement learning can be applied to directly guide models by non-differentiable metrics [Zhang *et al.*, 2019a], such as ROUGE. Zhao *et al.* [2020] utilized curriculum learning to let the model learn from easy documents to hard documents. Moreover, DIALOGPT [Zhang *et al.*, 2020] implemented a maximum mutual information (MMI) scoring function to alleviate generating bland, uninformative responses.

7 Conclusion and Future Outlooks

This paper presents an overview of the recent advances achieved in pretrained language models for text generation. We mainly summarize the extensions of PLMs in modeling different data types in input and satisfy special text properties in output. We also discussed several useful fine-tuning strategies for text generation.

To advance this field, there are several promising future directions for applying PLMs to text generation.

Model Extension. Although various extensions have been proposed in Section 3, there still exist discrepancies between pretraining and downstream generation tasks. For example, the "`[MASK]`" token in pretraining stage will not be used in fine-tuning stage, which further aggravates the pretraining-finetuning discrepancy. Thus, it further desires to design an appropriate pretraining paradigm for text generation. Moreover, incorporating external knowledge into PLMs during pretraining has been shown to be effective [Zhang *et al.*, 2019c], and it is promising to investigate how to inject related knowledge for text generation.

Controllable Generation. Controllable text generation with PLMs is an interesting direction but still at a very early stage. Controlling some attributes of the generated text has many useful applications such as generating positive response to patients with depression in dialogue systems. However, PLMs are usually pretrained in universal corpus, which is difficult to control the multi-grained attributes of the generated text (*e.g.,* sentiment, topic, and coherence). Keskar *et al.* [2019] has explored text generation with control codes that govern style, content and task-specific behavior. While, these control codes are preset and coarse-grained. Future work can explore multi-grained control and develop PLMs that are sufficiently steerable.

Model Compression. Although PLMs with large-scale parameters have achieved success in text generation, these models are challenging to be deployed in resource constrained environments. As a result, it is meaningful to study how to achieve competitive performance with a small number of parameters. Several methods have been proposed to compress PLMs, such as parameter sharing [Lan *et al.*, 2020] and knowledge distillation [Sanh *et al.*, 2019], whereas most of them focused on BERT-based models, and little attention has been paid to compressing PLMs for text generation.

Fine-tuning Exploration. The direct intention of pretraining is to distill the linguistic knowledge learned in PLMs to downstream generation tasks. And, fine-tuning is the predominant transfer method at present. There could be various ways to transfer knowledge from PLMs to downstream models. For example, Chen *et al.* [2020a] exploited knowledge distillation by adopting BERT as teacher model and a vanilla RNN generation model as student model. Through this method, the linguistic knowledge of BERT can be distilled into the downstream model.

Language-agnostic PLMs. Nowadays, almost all the PLMs for text generation are mainly based on English. These PLMs will encounter challenges when dealing with non-English generation tasks. Therefore, language-agnostic PLMs are worthy to be investigated, which need to capture universal linguistic and semantic features across different languages. An interesting direction is how to reuse existing English-based PLMs for text generation in non-English languages.

Ethical Concern. Currently, PLMs are pretrained on large-scale corpus crawled from the web without fine-grained filtering, potentially causing ethical issues such as generating private content about users. Therefore, researchers should try their best to prevent misusing PLMs. For this purpose, we can follow the key steps provided by Ross [2012], such as identifying threats and potential impacts and assessing likelihood. Besides, the text generated by PLMs might be prejudiced, which is in line with the bias in training data along the dimensions of gender, race, and religion [Brown *et al.*, 2020]. Hence, we ought to intervene PLMs for preventing such biases. The research on the general approach is extensive but still preliminary for PLMs.

Acknowledgments

This work was partially supported by the National Key R&D Program of China under Grant No. 2020AAA0105200, National Natural Science Foundation of China under Grant No. 61872369 and 61832017, Beijing Academy of Artificial Intelligence (BAAI) under Grant No. BAAI2020ZJ0301, Beijing Outstanding Young Scientist Program under Grant No. BJJWZYJH012019100020098, the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China under Grant No.18XNLG22 and 19XNQ047. Xin Zhao is the corresponding author.

References

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[Bao *et al.*, 2020] Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. PLATO-2: towards building an open-domain chatbot via curriculum learning. *arXiv preprint arXiv:2006.16779*, 2020.

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, and Nick Ryder et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[Chen *et al.*, 2020a] Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. Distilling knowledge learned in BERT for text generation. In *ACL*, 2020.

[Chen *et al.*, 2020b] Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang. Few-shot NLG with pre-trained language model. In *ACL*, 2020.

[Conneau and Lample, 2019] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *NeurIPS*, 2019.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[Dong *et al.*, 2019] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *NeurIPS*, 2019.

[Fan *et al.*, 2019] Zhiyun Fan, Shiyu Zhou, and Bo Xu. Unsupervised pre-training for sequence to sequence speech recognition. *CoRR*, arXiv preprint arXiv:1910.12418, 2019.

[Gehring *et al.*, 2017] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional sequence to sequence learning. In *ICML*, 2017.

[Gong *et al.*, 2020] Heng Gong, Yawei Sun, Xiaocheng Feng, Bing Qin, Wei Bi, Xiaojiang Liu, and Ting Liu. Tablegpt: Few-shot table-to-text generation with table structure reconstruction and content matching. In *COLING*, 2020.

[Gu *et al.*, 2020] Jing Gu, Qingyang Wu, Chongruo Wu, Weiyan Shi, and Zhou Yu. A tailored pre-training model for task-oriented dialog generation. *arXiv preprint arXiv:2004.13835*, 2020.

[Guan *et al.*, 2020] Wang Guan, Ivan Smetannikov, and Man Tianxing. Survey on automatic text summarization and transformer models applicability. In *CCRIS*, 2020.

[Hendrycks *et al.*, 2020] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. In *ACL*, 2020.

[Keskar *et al.*, 2019] Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. CTRL: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

[Kryscinski *et al.*, 2018] Wojciech Kryscinski, Romain Paulus, Caiming Xiong, and Richard Socher. Improving abstraction in text summarization. In *EMNLP*, 2018.

[Lan *et al.*, 2020] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *ICLR*, 2020.

[Lewis *et al.*, 2020] Mike Lewis, Yinhan Liu, and Naman Goyal et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020.

[Li *et al.*, 2019] Junyi Li, Wayne Xin Zhao, Ji-Rong Wen, and Yang Song. Generating long and informative reviews with aspect-aware coarse-to-fine decoding. In *ACL*, pages 1969–1979, 2019.

[Li *et al.*, 2020] Junyi Li, Siqing Li, Wayne Xin Zhao, Gaole He, Zhicheng Wei, Nicholas Jing Yuan, and Ji-Rong Wen. Knowledge-enhanced personalized review generation with capsule graph neural network. In *CIKM*, pages 735–744, 2020.

[Li *et al.*, 2021a] Junyi Li, Tianyi Tang, Gaole He, Jinhao Jiang, Xiaoxuan Hu, Puzhao Xie, Zhipeng Chen, Zhuohao Yu, Wayne Xin Zhao, and Ji-Rong Wen. TextBox: A unified, modularized, and extensible framework for text generation. In *ACL*, 2021.

[Li *et al.*, 2021b] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Zhicheng Wei, Nicholas Jing Yuan, and Ji-Rong Wen. Few-shot knowledge graph-to-text generation with pre-trained language models. In *Findings of ACL*, 2021.

[Li *et al.*, 2021c] Junyi Li, Wayne Xin Zhao, Zhicheng Wei, Nicholas Jing Yuan, and Ji-Rong Wen. Knowledge-based review generation by coherence enhanced text planning. In *SIGIR*, 2021.

[Lin *et al.*, 2020] Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. Pre-training multilingual neural machine translation by leveraging alignment information. In *EMNLP*, 2020.

[Liu and Lapata, 2019] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In *EMNLP*, 2019.

[Mager *et al.*, 2020] Manuel Mager, Ramón Fernandez Astudillo, Tahira Naseem, Md. Arafat Sultan, Young-Suk Lee, Radu Florian, and Salim Roukos. Gpt-too: A language-model-first approach for amr-to-text generation. In *ACL*, 2020.

[Peters *et al.*, 2018] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL-HLT*, 2018.

[Qiu *et al.*, 2020] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *arXiv preprint arXiv:2003.08271*, 2020.

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

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020.

[Ribeiro *et al.*, 2020] Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. Investigating pretrained language models for graph-to-text generation. *arXiv preprint arXiv:2007.08426*, 2020.

[Ross, 2012] R.S. Ross. Guide for conducting risk assessments. In *NIST Special Publication*, 2012.

[Rothe *et al.*, 2020] Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. Leveraging pre-trained checkpoints for sequence generation tasks. *TACL*, 2020.

[Sanh *et al.*, 2019] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[See *et al.*, 2017] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *ACL*, 2017.

[Song *et al.*, 2019] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: masked sequence to sequence pre-training for language generation. In *ICML*, 2019.

[Sun *et al.*, 2019a] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*, 2019.

[Sun *et al.*, 2019b] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[Wada and Iwata, 2018] Takashi Wada and Tomoharu Iwata. Unsupervised cross-lingual word embedding by multilingual neural language models. *arXiv preprint arXiv:1809.02306*, 2018.

[Wolf *et al.*, 2019] Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*, 2019.

[Xia *et al.*, 2020] Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, Xin Liu, and Ming Zhou. XGPT: cross-modal generative pre-training for image captioning. *arXiv preprint arXiv:2003.01473*, 2020.

[Xu *et al.*, 2020a] Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. Discourse-aware neural extractive text summarization. In *ACL*, 2020.

[Xu *et al.*, 2020b] Shusheng Xu, Xingxing Zhang, Yi Wu, Furu Wei, and Ming Zhou. Unsupervised extractive summarization by pre-training hierarchical transformers. In *EMNLP*, 2020.

[Yang *et al.*, 2020a] Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. CSP: code-switching pre-training for neural machine translation. In *EMNLP*, 2020.

[Yang *et al.*, 2020b] Ziyi Yang, Chenguang Zhu, Robert Gmyr, Michael Zeng, Xuedong Huang, and Eric Darve. TED: A pretrained unsupervised summarization model with theme modeling and denoising. In *EMNLP (Findings)*, 2020.

[Zaib *et al.*, 2020] Munazza Zaib, Quan Z. Sheng, and Wei Emma Zhang. A short survey of pre-trained language models for conversational AI-A new age in NLP. In *ACSW*, 2020.

[Zeng and Nie, 2020] Yan Zeng and Jian-Yun Nie. Generalized conditioned dialogue generation based on pre-trained language model. *arXiv preprint arXiv:2010.11140*, 2020.

[Zhang *et al.*, 2019a] Haoyu Zhang, Jingjing Cai, Jianjun Xu, and Ji Wang. Pretraining-based natural language generation for text summarization. In *CoNLL*, 2019.

[Zhang *et al.*, 2019b] Xingxing Zhang, Furu Wei, and Ming Zhou. HIBERT: document level pre-training of hierarchical bidirectional transformers for document summarization. In *ACL*, 2019.

[Zhang *et al.*, 2019c] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: enhanced language representation with informative entities. In *ACL*, 2019.

[Zhang *et al.*, 2020] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *ACL*, 2020.

[Zhao *et al.*, 2020] Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. Knowledge-grounded dialogue generation with pre-trained language models. In *EMNLP*, 2020.

[Zheng and Lapata, 2019] Hao Zheng and Mirella Lapata. Sentence centrality revisited for unsupervised summarization. In *ACL*, 2019.

[Zhou *et al.*, 2020] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *AAAI*, 2020.