文章编号：

可控文本生成技术研究综述

王舰 孙宇清

（山东大学 软件学院， 山东 济南 250101）

摘要：可控文本生成任务是指生成符合语法规则和语义需求，且满足给定约束的自然语言文本，具有重要应用价值。如何将约束嵌入到隐空间，从而有效控制离散的词汇生成过程是困难问题，特别是在复杂应用场景中，不仅需要控制文本内容，还要求形式多样、语言灵活以及生成的长文本逻辑合理等，这使得可控文本生成任务更具挑战且难以评估。近年来，数据驱动的神经方法得到广泛应用，特别是大规模预训练语言模型大幅度提升了生成文本质量。该文综述这些生成方法中的代表性技术架构和模型，讨论文本生成领域定性和定量评价指标，以及相关数据集；针对可控文本生成任务的文本多样性和句子间语义一致性等高层次需求，重点讨论相关技术前沿进展，分析其理论依据和技术优势。最后总结可控文本生成任务仍然面临的挑战和未来发展方向。

关键词：可控文本生成；文本评估；文本多样性；长文本生成

中图分类号：TP391　　　　**文献标识码：**A

Survey on Controllable Text Generation

WANG Jian, SUN Yuqing

(School of Software, Shandong University, Jinan, Shandong 250101, China)

Abstract: The controllable text generation task is to generate a natural language text that satisfies grammatical rules and semantic requirements under constraints, which is widely used in practical scenarios. It is difficult to embed the constraints into latent space to control the text generation process in an explicit way. Especially in the complex scenarios, the generated texts should be linguistically diverse and semantic consistency in addition to satisfying the constraints, which makes the controlled text generation more challenging. In recent years, the data-driven controllable text generation methods have become the mainstream approaches, especially the use of large-scale pre-trained language models and the generative adversarial networks further significantly improves the quality of generated text. We summarize the representative technical architecture and models, as well as the task-related datasets. Focusing on some challenging requirements such as the linguistic diversity and semantic relevance in long texts, we survey the theories and techniques of the related methods, as well as discuss the advantages and shortcomings.

Text evaluation is an important part of the text generation task. Most of the metrics evaluate the generated texts by the degree of word sequence matching with references. Since experiments demonstrate that these metrics are weakly correlated to the human evaluation, some researches introduce neural method to mimic human evaluation. We discuss these qualitative and quantitative metrics from the perspective of cognitive linguistics and present their merits and shortcomings. At last, we summarize the remaining challenges and present some promising research directions for the controllable text generation and evaluation.

Key words: controlled text generation; text evaluation; text diversity; long text generation

收稿收稿日期：20XX- XX-XX；**定稿日期：**20XX- XX-XX

基金项目：国家自然科学基金项目(62376138)；山东省自然科学基金创新发展联合基金重点项目(ZR2022LZH007)

0 前言

可控文本生成任务是指生成满足主题、情感和风格等约束、符合语法规则的自然语言文本，是人机交互的重要方式。早期使用基于模板的方法，只能控制文本结构，且生成文本的表达缺少多样性。近年来，随着文本生成的应用场景不断复杂，对于文本各方面的可控性需求也逐步增加。为了满足灵活多样的可控性需求，数据驱动的神经方法成为可控文本生成的基础方法，代表性的大规模预训练语言模型通过精调、提示学习等不同策略[1]生成满足约束的流畅文本。

从应用需求角度，可控文本生成任务除了生成满足给定约束的文本，在很多场景中，还需要关注生成文本的形式多样性和长文本生成时句子间的语义一致性两个高层次目标。为此，**可控多样性文本生成任务**是在属性可控前提下，生成多个语义接近但形式不同的文本。如情感可控的评论生成或者问答任务[2]。**可控长文本生成任务**是在属性约束下，生成语义与上下文一致、逻辑合理的长文本。如文本创作任务[3][4]。

本文第 1 部分介绍可控文本生成任务所涉及的基本知识，定义可控文本生成任务，并讨论用于文本生成的基础模型。第 2 部分综述可控文本生成任务相关的前沿进展，讨论技术细节和代表性工作。对于生成文本的形式多样性和句子间的语义一致性两个高层次目标，第 3、4 部分分别综述代表性方法，分析这些方法的理论依据和核心差异。

文本质量评估是衡量可控文本生成模型的重要组成部分。第 5 部分讨论评估指标，对比不同指标在评估粒度和核心思想上的差异，讨论这些指标在评估文本语义和表达方面的局限，并综述可控文本生成任务的标记数据集。

由于语言所承载的思想传播、情绪表达等复杂社会和文化功能，可控文本生成技术仍然存在很多挑战，如细粒度情感控制问题、文本幻觉问题、文本偏见问题等。第 6 部分将结合认知理论，从知识引导和数据驱动的生成方法、生成文本的可控性评价、细粒度属性控制等方面展开讨论，探讨未来发展方向。

1 基本知识

1.1 基础概念

可控文本生成任务期望生成的文本符合给定的属性，其生成过程形式化为：

$$p(y) = \sum_{t=1}^{r} p(y_t|y_{<t}, x, c) \tag{1}$$

其中，x为输入文本，y为生成文本，$p(y)$为生成y的概率，y_t为第t生成步生成的词，$y_{<t}$为前t步已生成的词序列，$p(y_t|y_{<t}, x, c)$表示 y_t的生成概率，r为生成文本长度，c为要控制的文本属性，包括关键实体[5]、情感极性[2]、主题[6][7]、风格[8]、人物角色[9]等。

1.2 基础模型

可控文本生成方法通常使用编码器解码器框架建模属性约束下输入文本到生成文本的映射过程，针对任务特点，编码器和解码器可以选择不同的网络结构。目前，大规模预训练模型通常会作为文本编码器或者解码器的首选[10][11] [12]。这些模型基于 Transformer 框架[13]，以语言建模为基础任务，在海量语料库上，学习到语言学知识和常识知识，从而具备语言理解、文本生成、类比推理等能力[14][15]，代表性模型包括 ChatGPT[16]，BART(Denoising Sequence-to-Sequence Pre-training for Natural Language Generation)[17]，GPT(Generative Pre-Training Transformer) 系列[18][19]，T5(Transfer Text-to-Text Transformer)[20]等。

为了使生成的文本符合用户偏好，相关方法在模型训练中引入强化学习方法 RLHF(Reinforcement Learning From Human Feedback)[21][22][23]，借助人工标注数据训练奖励模型评估生成文本，基于反馈信息使用策略梯度算法(Policy Gradient，PG)[24][25]、近端策略优化算法(Proximal Policy Optimization，PPO)[26]等更新预训练模型。近期具备较大影响力的 ChatGPT 通过多种预训练任务和 RLHF 微调技术训练模型，成为一个可用于多种任务的自然语言处理模型。GPT-4 模型[19]利用更多的人类反馈信息，包括一些安全性反馈，在逻辑推理、安全性等方面超越了 ChatGPT。

在可控文本生成领域，预训练模型已经成为文本生成任务的骨干网络，通过任务相关的训练策略、解码策略、与对抗文本生成框架、变分自

编码器结合等方式，实现对生成文本的控制。

2 可控文本生成方法

2.1 基于预训练模型精调的可控文本生成

这种方式以预训练模型中的语言知识为基础，通过在带有属性信息的数据集上精调，将属性信息融入文本生成过程。代表性工作如可插拔的语言模型(Plug-and-Play Language Model, PPLM)[27], 它在预训练模型基础上添加属性判别模型。在每一个生成步，利用属性判别模型提供的梯度信息微调预训练模型的隐藏层参数，当有新的属性需要控制时，可以添加新的属性判别模型，而无需重新训练整个语言模型。CoCon(Content-Conditioner)模型[28]在GPT-2模型的解码层中添加了内容控制层，该控制层通过自注意力机制增强属性和文本生成过程的关联，并通过多个自监督的损失函数训练。

2.2 基于提示学习的可控文本生成

该类方法通过构造和属性相关的提示，激发预训练模型的知识，从而在少样本或者无样本的场景下实现可控文本生成。基于提示学习的可控文本生成方法依据提示的不同形式，划分为两类，一是以离散词汇构成的可理解的指令作为提示的方法，另一类是以可训练的参数矩阵作为提示的方法。

指令方法通常给出可理解的任务描述，如“写一段文本，包含关键词：西藏、浪漫”，通过少量监督样本训练模型，使模型理解指令意图，随后用于未见指令或者未见任务中。后续方法通过引入多个任务、对同一个任务使用多个指令、引入思维链等方式进一步提升模型执行指令的能力。如FLAN(Finetuned Language Models via Instructions)[29]和T0(Multitask Prompted Training Based on T5)[30]使用多任务多指令的方式训练模型，随后在未见任务测试模型性能，证明了这种基于指令的方法在未见任务上的有效性。InstructGPT(Instruction GPT)[21]方法引入了基于人工反馈的强化学习方法RLHF，进一步提升模型执行指令的能力。STRUCTCTG(Controlled Text Generation Framework That Incorporates Instructions)[31]直接将需要控制的属性转为指令，构造多个指令-文本样本对，使得用户可以灵活的以自然语言的形式输入指令，从而控制文本生成。这些方法存在指令对文本控制能力逐步减弱的问题，即随着生成文本长度的增加，文本内容逐步不受指令控制。

针对该问题，IP(Inverse Prompting)方法[32]在生成过程中引入反向预测指令的策略，即在每个生成步，利用Beam Search策略形成多个生成序列，保留和指令相关程度高的序列进行后续生成。

另一类基于参数矩阵的方法通过监督数据优化提示在语义空间的表示，减少了人工设计提示的开销[33]。Prefix-tuning方法[34]使用一组随机初始化的前缀矩阵作为指令以表示要控制的属性，前缀和输入文本拼接作为预训练模型的输入，在训练过程中，预训练模型参数保持不变，只优化前缀对应的网络中每一层的参数。相似的，P-tuning方法[35]以对自然语言编码后的嵌入作为前缀，并插入部分核心词汇使其更具代表性。Multi-Control[36]方法的前缀来自从多个属性空间的交集中采样的向量，用于生成满足多个属性的文本。

为了更好的优化前缀矩阵，Prefix-NLG[33]引入属性间的对比损失L_d，从而有效利用属性间关系，如积极情感和消极情感的对立关系。如下式所示，其中，c为y的属性标签，c'为和c对立的属性标签，模型在训练过程中提升c下生成y的概率，同时降低c'下生成y的概率：

$$L_d = -log\frac{p(c)p(y|c,x)}{\sum_{c'} p(c')p(y|c',x)} \tag{2}$$

DisCup(Discriminator Cooperative Unlikelihood Prompt-tuning)方法[37]引入判别器判定已生成文本和属性的相关性，鼓励模型生成满足期望属性的词汇，同时减小和目标属性不一致的词汇出现概率。

2.3 基于条件预训练语言模型的方法

预训练模型在预训练任务中缺少属性信息的融入，导致应用到可控文本生成时缺少通用性。针对上述问题，最直接的方法是预训练一个条件语言模型，从而得到用于可控文本生成的基础大模型。这种方式提升了预训练模型在可控文本生成领域的适用性，是一个重要的发展趋势，最具代表性的工作为CTRL(Conditional Transformer Language Model For Controllable Generation)[38],

该模型将在大规模语料中和文本共现的属性如主题、领域代码、URL 等作为控制属性，直接训练得到带有属性信息的条件语言模型。在实际应用中，CTRL 模型可以作为一个基础的模型，为训练针对特定任务的可控文本生成模型提供高的起点。

2.4 属性引导的解码策略

该类方法引入属性判别器$D_c(c|x,y_{<t},y_t)$，基于输入文本x和已生成文本$y_{<t}$判定当前生成步的候选词汇y_t和属性c的相关性，利用此相关性，优化预训练模型生成的词汇分布而不更改预训练模型参数，从而在保证语言流畅度的同时，实现文本可控性。如在生成积极情感的评论时，对于已生成文本“这部电影很_ _”,依据预训练模型得到的词汇分布，后续词可能会选择“精彩”或者“枯燥”等。在引入属性判别器后，辅助模型判断‘精彩’这个词和积极情感相关程度更大，可以提升“精彩”的采样概率。

属性判别器有不同形式，如可预训练的神经网络或者是一个相关性度量函数。该判别器和预训练模型输出的词汇分布的结合方式也有不同。如 FUDGE(Future Discriminators for Generation)[39]方法按照如下式建立判别器判定结果和最终的词汇采样分布的关联：

$$P(y_t|x,y_{<t},c) \propto P_{lm}(y_t|x,y_{<t})D_c(c|x,y_{<t},y_t)^{\gamma} \tag{3}$$

其中，γ为权重系数，表示属性判别器对预训练模型输出词汇分布的影响强度。PPCG(Plug-and-Play Method for Controlled Text Generation)[40]方法使用相关性度量函数$S()$计算y_t和c的相关程度，累加到预训练模型输出的分布上：

$$P(y_t|x,y_{<t},c) = P_{lm}(y_t|y_{<t},x) + \lambda max\,(0,S(y_t,c)) \tag{4}$$

上述方法在每个生成步都施加属性约束，实际上，并不是每个生成词都要和属性相关，比如一些增强语言流畅度的连接词。因此，有必要在使用属性判别器前，确定当前位置是否应该生成和属性相关的词，从而避免影响语言流畅度。CAT-PAW(Controllable Text Generation with Position-Aware Weight)方法[41]在公式(3)中额外引入一个约束因子$f(c,y_{<t},y_t)$约束属性判别器的影响程度：

$$P(y_t|x,y_{<t},c) \propto P_{lm}(y_t|x,y_{<t})P_c(c|x,y_{<t},y_t)^{\gamma * f(c,y_{<t},y_t)} \tag{5}$$

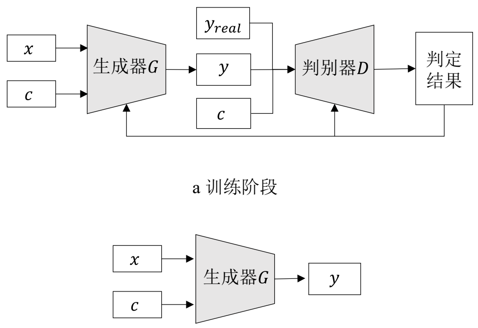
其中，约束因子$f(c,y_{<t},y_t)$的值为依据已生成序列预测当前生成步应该生成属性相关词的概率，其值越大，越倾向于生成属性词，它可通过多种方式得到，如在监督语料上训练的神经网络。

2.5 基于对抗生成结构的可控文本生成

对抗生成结构[42]引入了神经判别器评估生成文本和属性的一致性，进而指导文本生成。该结构包含生成器 G 和判别器 D 两个模块，判别器判定输入文本是否和真实文本的特征一致，并将判定结果反馈生成器，生成器目标是生成使判别器无法分辨真假的样本，判别器和生成器对抗训练优化。应用于可控文本生成任务中，生成器以x和属性 c 为输入，输出y，判别器判定文本y和属性 c 的一致性，为生成器提供反馈信息，其生成过程如图 1 所示，整体的优化目标为[43][44]：

$$\min_{G}\max_{D} L(G,D) = E_{y_{real}\sim P_{data}}[logD(y_{real},c)] + E_{y\sim G(x,c)}[log(1-D(y,c))] \tag{6}$$

其中，P_{data}为监督样本分布，$D(y,c)$为判别器判定文本y和属性c一致的概率。与上节所述属性引导的解码策略中所使用的属性判别器不同，该结构中的判别器参与模型的训练过程，和生成器以对抗方式逐步优化。



a 训练阶段

b 生成阶段

图 1 生成对抗网络在可控文本生成时的训练和生成过程

基于对抗框架，SentiGAN[2]使用多个生成器和一个多类别判别器，每个生成器负责生成一类属性的文本，提升了生成结果的准确性，判别器负责判定生成文本符合每个属性的概率，进而指导生成器。在多属性约束下的文本生成场景中，上述方法中的判别器的类型或者数量可以灵活增

加，如 PT-GAN(Professional Text Generative Adversarial Networks)模型[45]同时使用了语言连贯性判别器、专业领域相关性判别器等控制文本内容和文本特征，同时实现了对多个不同属性的控制。

对抗生成结构存在连续空间的优化过程和在离散词汇空间采样的矛盾[46]，即词汇采样过程不可导，判别器对生成文本的判定结果无法通过梯度反向传播。针对上述问题，有两类解决方法：

一是使用连续可导的函数近似不连续的采样过程。如使用连续的 *Gumbel-softmax* 函数[47]作为采样操作的近似[48]，在第t个生成步，将模型前向过程预测的概率分布π_t和 *Gumbel* 分布累加作为词汇概率分布O_t，然后从O_t中选择最大概率的词：

$$O_t = softmax((\delta + log\,\pi_t)/\tau) \tag{7}$$

其中，随机变量$\delta \sim Gumbel(0,1)$，τ用于调节采样倾向。前向生成过程在概率分布O_t下采样词汇，反向传播过程通过上式计算梯度。引入 *Gumbel* 分布的另一个优点是从O_t中以最大概率采样的结果近似于从π_t中按照概率采样的结果，这比依据π_t直接以最大概率采样更为合理。

另一方法是借助强化学习[49]，将文本的生成过程看作是马尔可夫决策过程，状态Z_t为前 t-1 步生成的文本$y_{<t}$，动作A_t为生成一个词y_t，状态Z_t下执行动作A_t获得的奖励$Q(Z_t, A_t)$由判别器反馈。生成器以最大化生成文本的期望奖励为目标，使用策略梯度的方法更新[50]，生成器的梯度为：

$$\nabla_G = \textstyle\sum_{t=1}^{r} Q(Z_t, A_t) \nabla \log p_G(y_t|x, y_{<t}) \tag{8}$$

其中，$p_G(y_t|x, y_{<t})$为生成模型生成y_t的概率。奖励 $Q(Z_t, A_t)$ 有不同的计算方法，如 SeqGAN(Sequence Generative Adversarial Nets)模型[51]以对$y_{<t}$补全后的文本所获得的整体奖励作为当前状态的奖励$Q(Z_t, A_t)$，MaskGAN(Mask Generative Adversarial Nets)[52]方法中的奖励$Q(Z_t, A_t)$为判别器判定词汇y_t为正确词汇的概率。

综上，针对可控性文本生成问题，我们概述了近期代表性方法的核心思想、优点和不足，并给出了需要进一步优化的方向，如表 1 所示。

3 可控多样性文本生成

可控多样性文本生成旨在针对一个输入文本，生成多个满足给定属性、内容丰富的文本，如在评论生成任务中生成多个具有积极情感、内容丰富的商品评论，避免出现“好”、“我同意”等合乎控制属性但无意义或者重复的文本。当前的评估方法、词汇的采样策略等限制了模型的多样性文本生成能力，具体体现在一是缺少文本多样性评估指标，现有评估方法如 Self-BLEU[53]、Distinct-n[54]等基于生成文本词序列的用词差异性，无法体现文本表达的灵活性[55][56]；二是生成过程中使用的最大概率的采样策略导致模型选择无实际意义的高频词汇，减弱文本内容丰富程度。针对上述问题，相关方法从采样策略、模型结构、评估等角度探索以增强文本多样性。

3.1 基于采样策略的方法

该类方法在词汇采样阶段引入一定的随机性，以探索自然语言不同的表达模式，从而生成多个有差异性的文本[59]。如波束搜索(Beam Search，BS)方法在每个生成步维持 B 个高概率候选序列，每个候选序列包含不同的用词。但是，实际上，B 个序列之间差异性很小[57]。DBS (Diverse Beam Search)[58]将多个生成序列分组，组内序列的生成和普通的波束搜索方法一致，但鼓励组间语义差异性，使每个组向不同方向探索。

上述方法本质是选择具有高概率的多个序列，这使得这些序列在用词上较为保守，不符合人类用词习惯，甚至会出现重复用词现象。为此，Top-K Sampling 方法[60]在每个生成步，从预测的词汇概率分布中选择前 K 个最高概率的词作为候选，概率归一化后从这些词中采样，提升用词多样性。但是，K 是固定的，在词汇概率分布平滑即词汇概率相差不大时，容易漏掉具有较高概率的合理词，在词汇分布陡峭时，容易引入低概率的无关词。Nucleus Sampling 方法[61]设置一个阈值，在每个生成步从词汇表中选择总概率达到阈值的最小候选词集，在此候选词集中采样，从而降低引入低概率无关词的可能性。

针对生成文本用词重复问题，一类方法在采样阶段显式约束当前生成词，如 Welleck 等人在模型训练过程中，降低已经出现在生成文本中的词的概率[62]，Contrastive Search 方法[63]计算生成词和已生成的每个词的语义相似度，规避语义相似度高的词汇。另一类是基于判别器的方法，如 DP-GAN(Diversity-Promoting GAN)模型[64]，该模

表 1 可控文本生成方法对比

分类	方法	核心思想	优点	不足	优化方向
基于精调的方法	PPLM[27]	添加一个属性模型纠正语言模型	使用方式灵活	解码策略复杂，文本生成速率较低	深度挖掘预训练模型的知识；句法、篇章结构等文本核心要素的控制；
	CoCon[28]	通过自监督的目标函数提升生成文本对条件的依赖	以文本为控制变量，实用性强	由于文本变量的高自由度，需要大量训练语料	
基于提示学习的方法	InstructGPT[21]	引入 RLHF 技术，提升模型遵循指令的能力	生成的文本更符合人类偏好	训练开销较大	提升对复杂指令的理解能力
	STRUCTCTG[31]	构造和属性相关的指令	训练所用指令和属性更相关，控制能力强	对未见属性缺少鲁棒性	
	Prefix-NLG[33]	引入属性间的对立关系优化前缀	对立属性的引入提升了属性对文本的控制能力	如果属性间不存在对立关系(如长度和情感)则无法适用	指令的优化策略；提升生成过程对指令的依赖
	DisCup[37]				
	Multi-Control[36]	多个属性空间的重合部分对应满足多属性的文本	满足多属性控制场景	为了构建属性空间，需要较多的训练语料	
	IP[32]	利用生成文本反向预测前缀	提升了前缀对文本的控制能力	反向预测时，需要人为变换前缀的格式，使预测方式符合常理	
条件预训练模型	CTRL[38]	使用带有属性信息的文本预训练	具有通用性	训练代价非常大	减小计算开销
属性引导的解码策略	PPCG[40]	限定生成词汇和属性相关词的语义距离	使用方式灵活	在每个生成步都施加约束，影响文本流畅性；计算开销大	约束词汇生成的同时保持文本语言流畅性；同时约束多个生成步，而非独立的约束每个生成步
	FUDGE[39]	训练属性判别器模型修正预训练模型输出的词汇分布	满足多属性控制场景		
	CAT-PAW[41]	先判定是否应该生成和属性相关的词，再调整词汇分布	在每个生成步动态的调节辅助模型的影响，提升了生成文本的语言流畅性	基于词汇级别的判定信息，没有考虑全局语义；不适用于风格等无法用词袋表达的属性	
基于对抗的方法	CTGAN[44]	利用判别器提供生成文本和属性的一致性反馈	方法简洁，易于使用	条件变量和生成过程缺少交互，变量控制能力较弱	训练高质量的判别器
	SentiGAN[2]	多生成器、一个多类别判别器结构	每个生成器生成一类文本，准确率高	模型结构和属性的数量相关，不够灵活	
	PT-GAN[45]	多生成器，多判别器结构	面向专业领域文本，同时实现多属性的控制	训练方法复杂，计算开销较大	

型使用预训练语言模型作为判别器，判定生成文本的出现概率，重复的词汇因为和正常语言的用词模式存在差异，会被赋予低的概率，由此，形成反馈优化生成模型[64]

3.2 基于变分自编码器的方法

变分自编码器(Variational Auto-Encoders，VAE)[65]基于编码器解码器结构，它假定文本x可以通过服从高斯分布的隐变量$z{\sim}N(0,I)$经过参数θ的神经网络映射得到，其中I为单位矩阵。VAE 通过最大化训练集中所有样本的生成概率优化参数θ：

$$P(x)=\int P(x|z;\theta)\,P(z)dz \tag{9}$$

真实的训练过程中，VAE 通过最大化$logP(x)$的一个变分下界最大化$P(x)$，表示为：

$$ELBO(x)=E_{z\sim Q(z|x)}[logP(x|z)]-D_{KL}[Q(z|x)||P(z)]\le logP(x) \tag{10}$$

上式中，$Q(z|x)=N(z|\mu(x),\sigma^2(x)I)$是编码器对$x$编码得到的隐变量后验分布，$\mu(x)$和$\sigma^2(x)$为该分布的均值和方差，$P(z)=N(0,I)$为假定的隐变量先验分布，$P(x|z)$为解码器依赖隐变量$z$重构输入文本的概率分布，$D_{KL}$表示两个分布的 KL 散度(Kullback-Leibler Divergence)。应用阶段，为了生成多样性的样本，可以从分布$P(z)$中随机的多次采样 z，输入解码器生成样本，如图 2 中实线所示。

针对更广泛的序列到序列的文本生成任务需求，条件变分自编码器(Conditional Variational Autoencoder，CVAE)[65][66]在变分自编码器的基础上，将输入文本x作为目标文本y的条件，编码和重构目标文本y，从而能够建模文本x到y的映射过程。训练和使用过程如图 2 中虚线所示，其变分下界$ELBO(y|x)$为:

$$ELBO(y|x) = E_{z\sim Q(z|x,y)}[logP(y|z,x)] - D_{KL}[Q(z|x,y)||P(z)] \quad (11)$$

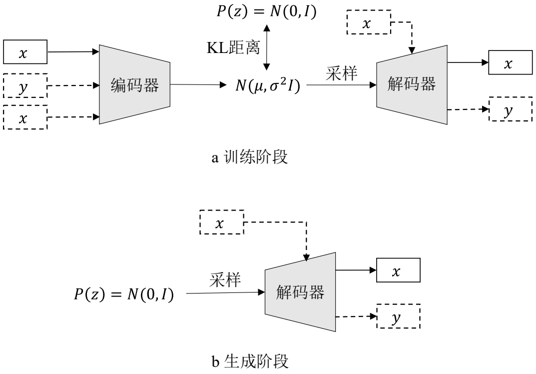


图 2 变分自编码器训练和生成过程

变分自编码器建立了有限维空间连续分布到文本空间的映射，在连续分布采样的隐变量是多个生成要素的融合，这为可控多样性文本生成提供了结构基础。基于变分自编码器框架，相关方法将决定文本生成的不同属性解耦到不同的隐空间，由此，在保持某一个属性对应的隐变量不变的情况下，通过从其他隐空间采样，实现文本可控性和多样性。

DSS-VAE(Disentangled Syntactic and Semantic Variational Autoencoder)模型[67]将文本编码为语义隐变量和句法隐变量，通过使用语义隐变量预测输入文本的词袋，使用句法隐变量预测句法结构，使两类隐变量分别包含语义和语法信息。基于离散隐空间的变分自编码模型 FVN (Focused-Variation Network)[68]使用包含多个向量的查询表作为隐变量的采样空间，查询表中每个向量代表一个聚类中心，使得模型能关注重要生成因素，忽略细节因素。SIVAE (Syntax-Infused Variational Autoencoder)模型[69]则引入了句法结构信息，分别编码文本和对应的句法树，更好的解耦和控制语义和句法。MixPoet(Mixed Latent Space for Diverse Poetry Generation)[70]方法通过引入对抗损失将影响古诗生成的因素，如背景、主题等编码入多个子空间，随后，在不同子空间采样生成不同的古诗。

基于变分自编码器的方法中在训练过程中存在后验塌陷问题[71]，即在训练初期隐变量后验分布$Q(z|x)$迅速接近先验分布$P(z)$，解码器退化成语言模型，不再依赖隐变量。针对该问题，相关工作有两类解决方法：一是提升隐变量同输入文本的关联程度，包括约束隐变量和输入文本的互信息[72][73]、引入最优隐变量[74]、差分不同特征文本所对应的隐空间[75]、在训练过程中动态调节 KL 散度的权重[71][76]、使用超球面上的 von Mises-Fisher 分布作为隐变量的分布[77]等。二是减小解码器对已生成文本的依赖范围，提升解码器对隐变量的依赖能力[71]，可采用 Dropout 机制[78]或随机将输入解码器的文本的部分词置为空，或以扩张卷积神经网络作为解码器[79]等方法。

3.3 基于扩散模型的方法

扩散模型(Diffusion Model)[80]是基于隐变量的生成模型，它通过建立属性和隐变量的关联，实现可控文本生成，通过隐变量采样的随机性，提升生成文本多样性。针对一个样本，该模型包括前向的加噪过程和反向的去噪过程。在加噪过程，模型对样本逐步添加已知噪声，每加一次噪声得到一个隐变量，直至最终的隐变量z_T为一个纯高斯噪声。在去噪过程，由去噪网络逐步去除z_T中的噪声，恢复原始样本。前向过程得到的隐变量作为监督信息训练去噪网络，训练后的去噪网络以随机采样的高斯噪声为基础，生成数据。扩散模型和 VAE 方法有相似性，但也有显著的不同，体现在扩散模型自定义一个由样本到高斯分布的过程，而非由网络学习这一过程。

扩散模型在去噪过程中需要预测多个隐变量，这使得我们可以多次融入属性信息。Diffusion-LM(Diffusion Language Model)[81]是其中的代表性方法，在去噪过程，它引入一个分类器判定隐变量和属性 c 的一致性。分类器在每个去噪步提供梯度信息指导生成和属性 c 相关的隐变量。区别于 Diffusion-LM 使用分类器的方式，LD4LG(Latent Diffusion for Language Generation)[82]直接将属性的嵌入和隐变量拼接，通过注意力机制融合两者信息指导去噪过程。

DIFFUSEQ(Diffusion for Sequence to Sequence)[83]模型拼接输入文本和目标文本，在训练过程保持输入文本不变，仅对目标文本加噪和去噪，用于序列到序列的生成任务。通过对隐变量的采样，DIFFUSEQ 能够生成多个高质量文本。

综上，针对可控多样性文本生成问题，本节综述了前沿代表性方法的核心思想、优点和不足。汇总如表 2。

表 2 可控多样性文本生成方法对比

分类	方法	核心技术	优点	不足
基于采样策略的方法	BS/DBS[58]	选择有高概率的多个序列分支	方法简单	倾向于选择高频常用词
	Top-K Sampling[60]	以概率排名前 K 的词作为候选词	避免了出现较多的高频常用词；非常灵活	对词汇分布敏感
	Nucleus Sampling[61]	以累积概率达到一定阈值的词作为候选词	降低了采样过程对词汇分布的敏感程度	阈值难以确定
	UL[62]	将对重复词的约束融入训练过程	使用高概率词同时避免重复用词	对重复词的惩罚权重难以确定；一些重复用词是合理的
	Contrastive Search[63]	在采样阶段约束重复用词		
	DP-GAN[64]	由判别器判定文本用词的合理性	判别器具有鲁棒性，能够包容一些重复用词行为	模型较难训练
基于变分自编码器的方法	DSS-VAE [67]	解耦句子的语义和句法	能在连续空间施加对属性的控制	不同属性对应的隐空间有重叠，减弱了多样性表达
	SIVAE[69]			
	FVN[68]	使用包含有限个向量的查询表作为隐变量取值空间	使模型更关注于核心生成要素	查询表中的向量数量不易确定
	MixPoet[70]	解耦古诗的多个生成因素到不同子空间	能够同时控制多个生成因素	子空间需要有交集且数量难以确定
基于扩散模型的方法	Diffusion-LM[81]	引入分类器控制生成过程的中间环节	适用复杂的可控场景	语言流畅性较低，解码速度较慢
	DIFFUSEQ[82]	在加噪过程保留输入文本语义，仅加噪目标文本	将扩散模型应用于序列到序列生成任务	
	LD4LG[82]	直接将属性编码和隐向量拼接	简单灵活	较难适用复杂控制场景，如句法可控的任务

4 可控长文本生成

在可控长文本生成过程中，随着文本长度的增加，生成的句子间容易出现逻辑不一致、语义漂移。人类在长文本写作中，依据全局语义逐步规划句子语义，随后，选择合适的词汇表达句子，并且写作过程中不断评估当前句子和整体语义及已完成句子的关联性。受此启发，前沿技术或是采用多步层次化的生成模式建模高层语义到底层写作的过程，或是通过判别器判定句间内容关联性，指导生成器，或者引入外部知识，使模型更好感知句子关联。

4.1 基于多步生成策略的方法

该类方法首先依据输入文本的语义生成高层的内容规划，随后依据内容规划生成具体的词汇。代表性的是层级编码器解码器 H-AE (Hierarchical Neural Autoencoder)[84]，编码阶段，H-AE 融合词的编码得到句子表示，依赖多个句子表示产生整个文档表示，解码阶段，模型在文档表示下规划每个生成句子的内容，以向量的形式表示，该规划向量作为解码器的初始状态生成句子。基于当前大规模语言模型的进展，HS-GPT(Hierarchical Structure and GPT-2)方法[85]沿用了上述结构，但引入预训练模型完成生成过程。PLANET 模型[86]利用自注意力机制动态关注每个已生成句子的规划向量和句子内容，由此计算下一个句子的规划向量，提升了规划的合理程度。上述方法在向量空间表达规划的内容，可以端到端的训练，但是规划缺少可解释性。

也有一些方法使用自然语言作为规划。相关工作首先规划全文的整体内容框架[87][88]或者语义演变序列[89][90]，然后填充细节内容。内容框架和语言演变序列可通过神经网络预测得到如 EP-PG(Event Transition Planning)模型[89]，也可通过外部知识图谱如 MKR(Multi-Level Knowledge Aware Reasoning)模型[90]得到。

4.2 基于语义一致性判别器的方法

为了有效地监督文本生成过程，基于判别器的模型或使用更细粒度的反馈形式，及时纠正生成文本时语义的漂移，或使用判别器反馈额外的内容逻辑关联性信息，约束文本的整体语义。

从反馈形式上，LeakGAN(GAN with Leaked Information)模型[91]每生成一个词，都会利用判别器提取当前已生成词序列的特征信息，然后泄露给生成器指导下一个词的生成。这种词汇级的反馈方式比在整个文本生成后才获得反馈的方式更为有效。

从反馈的内容上，RTT-GAN (Recurrent Topic-Transition Generative Adversarial Network)模型[92]针对新生成的句子，除了判定是否和真实文本特征一致，还引入主题判别器判定是否和已生成的句子主题相关，由此增强和已生成文本句子间的语义一致性。

4.3 基于关联性知识增强的方法

该类方法通过外部知识或者多任务学习方法来增强模型对句子关联关系的感知能力[93][84]，提升生成文本句子间的语义一致性。

Guan 等人通过自定义的模板将结构化知识库 ConceptNet[94]和 ATOMIC[95]中的关联信息转为自然语言文本，利用转换后的文本训练模型[96]。同时，通过扰乱句子顺序、替换某些句子、随机重复一些句子构造三类负样本训练模型，鼓励模型减少负样本生成概率，提升生成文本的合理性，形成如下损失函数：

$$L_{st} = L_{LM} + L_{CLS} \tag{12}$$

$$L_{LM} = -\log p(y_t|y_{<t}) \;\; y \in DA_1 \tag{13}$$

$$L_{CLS} = -\log p(c_y|y) \;\; y \in DA_1, DA_2, DA_3, DA_4 \tag{14}$$

其中，L_{LM}为文本预测损失函数，L_{CLS}为文本类别预测损失函数，DA_1表示正样本，DA_2, DA_3, DA_4表示三类负样本，c_y表示y所属的真实类别。

Guan 等人的另外一项工作 HINT(High-level Representations for Long Text Generation)[97]添加生成句子间的相似性预测任务优化句子语义表示，两个句子的真实相似性由预训练的 SentenceBert[98]标注，同时，HINT 利用句子顺序判别任务学习文章结构信息，使解码器获取高层语义特征，生成更具逻辑的文本。

针对可控长文本生成问题，本节综述了相关方法的核心思想、技术细节、优点和不足，汇总如表 3。

5 评估方法和数据集

基于数据驱动的神经文本生成方法通常不具备可解释性，那么判定模型质量的标准只能依赖对生成结果的评估。常用的评估指标依据有无参考文本，划分为两类，一类是针对有参考文本的情况，依据参考文本和生成文本的一致性进行评价，形式化为：给定生成文本$y = y_1 y_2 \dots y_r$和参考文本集合$Ref = \{y^1, \dots y^i, \dots, y^m\}$，从词序列的角度评价生成文本和参考文本$y^i$的匹配度，或从隐式空间评估两者的语义相似性。另一类方法是针对无参考文本的情况，该类方法通常是依据生成文本自身的统计特征或者外部知识来评估。本章综述这些文本评估方法，对比这些方法的设计思路和技术细节，讨论其优势和不足，并汇总相关数据集。

5.1 基于参考文本的评估

5.1.1 基于元组匹配的评估

n 元组是指连续 n 个词构成的序列，设$G_n(y)$表示生成文本 y 的所有 n 元组集合，$C(g,y)$表示元组g在文本 y 中的出现次数，$M_n(y, y^i) = G_n(y) \cap G_n(y^i)$表示生成文本 y 和参考文本y^i匹配的 n 元组集合。该类评估方法将文本看作元组集合，使用精确率衡量生成文本相对参考文本的内容匹配程度，使用召回率评估生成文本对参考文本的内容覆盖。

BLEU(Bilingual Evaluation Understudy)[100]方法兼顾元组的精确率p_n和文本长度度量生成文本质量，其计算公式如下：

$$BLEU(y, Ref) = BP * exp(\textstyle\sum_{n=1}^{N} W_n * log(p_n)) \tag{15}$$

$$BP = \begin{cases} 1 & r > \hat{r} \\ e^{(1-\hat{r}/r)} & r \le \hat{r} \end{cases} \tag{16}$$

$$p_n = \frac{\sum_{g \in G_n(y)} min\{C(g,y), \max_{1 \le i \le m} C(g, y^i)\}}{\sum_{g \in G_n(y)} C(g,y)} \tag{17}$$

其中，N 表示最大元组长度，权重W_n表示元组权重，通常取$1/N$，$\hat{r}$表示最接近生成文本长度r的参考文本长度，BP 为短文本惩罚系数，当生成文本过短时减小最终得分。

针对 BLEU 赋予每个元组相同权重，不能差别对待元组的问题，指标 NIST(National Institute of standards and Technology)[101]引入信息量区分不同元组的重要程度，综合元组精确率和文本长度评估文本质量：

$$NIST(y, Ref) = BP * \sum_{n=1}^{N} \left\{ \frac{\sum_{g \in M_n(y, y^i)} info(g)}{\sum_{g \in G_n(y)} C(g,y)} \right\} \tag{18}$$

$$BP = exp\,(\beta * (log^2(min\{r/\bar{r}, 1\}))) \tag{19}$$

$$info(g) = log_2 \left(\frac{\sum_{1 \le i \le m} C(g[0:-1], y^i)}{\sum_{1 \le i \le m} C(g, y^i)} \right) \tag{20}$$

其中，$\bar{r}$为参考文本的平均长度，β 是经验值常数，通常取 0.5，$g[0:-1]$表示元组g去除最后一个词后

表3 可控长文本生成方法对比

分类	方法	核心思想	优点	不足
多步生成	H-AE[84]	基于输入文本语义得到要生成的每个句子的规划向量，利用规划向量指导写作过程	整个过程是可导的，规划过程可以得到更新	模型结构复杂、注意力计算开销大
	HS-GPT[85]			规划过程未考虑已生成文本，导致错误累积[99]
	ProGen[88]			
	EP-PG[89]	学习一个内容规划模型生成规划	规划的内容具有可解释性	需要预先构造数据集训练规划模型；过程不可导
	MKR[90]	基于知识图谱构造事件规划序列		额外引入文本到图的任务，存在错误累积；过程不可导
	PLANET[86]	融合已生成文本语义动态规划写作内容	建模了长文本生成过程中生成文本语义的动态转换	解码过程较为复杂，计算开销大
基于判别器的方法	LeakGAN[91]	词汇级的信息反馈	增加了内容关联性反馈	逐词反馈开销较大
	RTT-GAN[92]	增加额外的主题一致性判别器		模型不容易优化和收敛
句子关联知识增强	KG-PM[96]	利用知识文本进一步训练模型	方法简单，通用性好	将三元组知识转为文本，有额外的错误累积和计算开销
	HINT[97]	通过句子关联知识，提升解码器对语义的感知能力		受限于外部句子关联性知识的正确性

构成的元组，信息量$info(g)$反映了元组g相对于参考文本的重要性，如果g完整出现的频次较小，而g中部分词汇出现的频次大，说明g更倾向于是一个低频组合词，应给予更多的权重。

ROUGE-n(Recall Oriented Understudy of Gisting Evaluation-ngram)[102]通过元组召回率度量生成文本覆盖参考文本内容的比例：

$$ROUGE-n = \frac{\sum_{g\in M_n(y,y^i)} max\{C(g,y),C(g,y^i)\}}{\sum_{g\in G_n(y^i)} C(g,y^i)} \tag{21}$$

ROUGE-L(Recall Oriented Understudy of Gisting Evaluation-Longest Common Subsequence)[102]综合元组精确率和召回率，从文本子序列匹配角度评估生成文本。一个词序列片段是文本的子序列当且仅当该片段中每个词出现在文本中，且每个词在文本中的索引值形成的整数序列是严格递增的。ROUGE-L 以匹配的最长子序列长度$L(y,y^i)$和生成文本长度 r 的比值作为精确率 P，以$L(y,y^i)$和参考文本长度r_i的比值作为召回率 R，并计算 F 值综合评估生成文本，具体计算公式如下，其中$\beta = P/R$。

$$P = L(y,y^i)/r \tag{22}$$

$$R = L(y,y^i)/r_i \tag{23}$$

$$F = \frac{(1+\beta^2)*P*R}{R+\beta^2*P} \tag{24}$$

METEOR(Metric for Machine Translation)[103]不仅考虑元组匹配数量和长度，而且鼓励位置连续共现元组，引入$chunks$表示生成文本和参考文本共现元组构成的匹配块集合，该集合的初始元素为生成文本中所有和参考文本匹配的1元组，当存在位置连续的匹配元组时，合并匹配元组为一个匹配块。METEOR 计算公式如下，其中，$Pen(y^i,y)$为对y的惩罚项，匹配块越多代表共现元组位置越不邻接，将给予一定的惩罚，P_i、R_i、F_i分别代表以y^i为参考文本的 1 元组的精确率、召回率及 F 值。

$$METEOR(y,Ref) = \max_{1\le i\le m}\{F_i*(1-Pen(y^i))\} \tag{25}$$

$$F_i = \frac{10*P_i*R_i}{R_i+9P_i} \tag{26}$$

$$P_i = \frac{\sum_{g\in M_1(y,y^i)} C(g,y)}{\sum_{g\in G_1(y)} C(g,y)} \tag{27}$$

$$R_i = \frac{\sum_{g\in M_1(y,y^i)} C(g,y)}{\sum_{g\in G_1(y^i)} C(g,y^i)} \tag{28}$$

$$Pen(y^i,y) = 0.5*\left(\frac{|chunks|}{\sum_{g\in M_1(y,y^i)} C(g,y)}\right)^3 \tag{29}$$

5.1.2 基于词序列编辑距离的评估方法

编辑距离指通过编辑操作将生成文本转换为参考文本的最小编辑次数或最小编辑代价。编辑操作包含词或字符的插入、删除、移动、替换等。TER (Translation Edit Rate)[104]以词为编辑对象，使用编辑距离和参考文本平均长度的比值计算文本相似性。CharacTER(Character Translation Edit Rate)[105]考虑同形词间的编辑距离应该小于非同

形词间的编辑距离，因此，以字符为编辑对象，采用编辑距离和生成文本长度的比值表示文本相似性。ITER(Improved Translation Edit Rate)方法[106]添加了同形词的匹配操作，计算方法如下：

$$\frac{min_{1\le i\le m}\{cost(y,y^i)\}}{r+k+min_{1\le i\le m}\{cost(y,y^i)\}} \tag{30}$$

其中，r表示为生成文本长度，k 为匹配的同形词数量，$cost(y,y^i)$表示生成文本到参考文本的编辑代价。

5.1.3 基于词序列语义的评估方法

语言的灵活性体现在使用不同词汇和形式的文本可以表达相同语义。为了包容这种特性，可以将文本转化为抽象的语义向量，例如向量均值法 EmbedAvg(Embedding Average)采用构成文本词向量的均值作为语义向量[107]，向量极值法 VecExt(Vector Extrema)采用文本中词向量的每一维的极值形成语义向量[108]，然后使用语义向量之间的余弦值评估生成文本和参考文本的语义一致性。

BERTScore(Bidirectional Encoder Representations from Transformer Score)[109]通过 BERT[110]得到y和y^i的词向量序列，利用如下公式计算生成文本词汇相对参考文本词汇的语义精确率P和召回率R，F 值，其中r，r_i分别为生成文本和参考文本的长度，w，v分别为词w，v的词向量。

$$P=\frac{1}{r}\sum_{w\in y} max_{v\in y^i}\{\mathrm{w}^\intercal \mathrm{v}\} \tag{31}$$

$$R=\frac{1}{r_i}\sum_{v\in y^i} max_{w\in y}\{\mathrm{w}^\intercal \mathrm{v}\} \tag{32}$$

$$F=\frac{2*P*R}{P+R} \tag{33}$$

WMD (Word Mover’s Distance)[111]以参考文本和生成文本间的最小词汇转移距离评价文本相似度。词汇转移距离是词频的转移量和词汇语义距离的累积。生成文本和参考文本分别使用归一化后的词频向量$d,d'\in R^{|V|}$表示，其中，$|V|$表示字典大小，$d[i]$和$d'[k]$分别表示生成文本的第 i 个词w_i和参考文本第 k 个词w'_k的词频，词汇语义距离$dis(w_i,w'_k)$为对应词向量w_i和w'_k的欧式距离。生成文本和参考文本的相似度为：

$$min_{A\ge 0}\sum_{i,k=1}^{|V|} A_{i,k}*dis(w_i,w'_k) \tag{34}$$

$$s.t.\ \sum_{k=1}^{|V|} A_{i,k}=d[i],\ \sum_{i=1}^{|V|} A_{i,k}=d'[k]$$

其中，$A\in R^{|V|*|V|}$是一个需要优化的词频转移矩阵，$A_{i,k}$表示生成文本第 i 个词到参考文本第 k 个词的词频转移量。为了区分词汇的重要程度，WRD(Word Rotator's Distance)方法[112]改进了 WMD，使用词向量w_i的范数归一化后作为$d[i]$，如下述公式，范数越大的词对句子语义的贡献越大，随后，使用词的余弦相似度度量词汇距离$dis(w_i,w'_k)=1-cos(w_i,w'_k)$。

$$d[i]=\frac{\|\mathrm{w}_i\|}{\sum_i\|\mathrm{w}_i\|} \tag{35}$$

WMD 和 WRD 方法从词汇角度度量文本相似性，当文本较长时，存在较大的计算开销。为了解决上述问题，SMS(Sentence Mover’s Similarity)[113]方法从句子级别度量文本相似性，将文本视作句子的组合，求解和 WMD 相似的优化问题。

5.1.4 神经评估方法

人工评估是最权威的评估方法，人类在评估文本时关注于文本整体的实用性、可读性，而非局限于单个的词汇。上述评估指标从词序列的角度机械化地比对生成文本和参考文本，没有从句子整体的角度评估文本，导致其评估结果和人工评估结果的关联性较弱[114][115]。因此，一些方法使用神经网络拟合人工评估结果。

APES(Answering Performance for Evaluation of Summaries)[116]基于参考文本中的命名实体生成多个问题，通过预训练的问答模型以生成文本为上下文回答问题，能被正确作答的问题比例反映生成文本的语义正确程度。相似的方法，QAEval(Question Answering Evaluation)[117]使用参考文本中更具代表性的名词短语来生成问题，提高问题的有效性。

BLEURT (Bilingual Evaluation Understudy with Representations from Transformers)[118]首先构造形如$(s,\tilde{s})$的多样化数据：s为维基百科的句子，$\tilde{s}$是对s扰动如删除某些词、掩码某些词等得到的句子。使用 BLEU, ROUGE-n 等自动化评估指标对$\tilde{s}$打分的结果预训练评估网络，提升网络鲁棒性。随后以人工对$\tilde{s}$打分结果为监督信息，精调评估网络，增加和人工评估的关联性。

BARTScore(BART for Evaluating Generated Text)[119]方法通过预训练的 BART 模型[17]计算输入文本x、生成文本 y 和参考文本y^i之间的转换概率，以此反应文本质量，如下式，其中$W(y_t)$表示t时刻生成的词汇权重。

$$BARTScore=\sum_{t=1}^{r} W(y_t)\,logp(y_t|y_{<t},x) \tag{36}$$

如果以$(y^i \rightarrow y)$组合作为 BART 的输入和输出，BARTScore 值反映了生成文本的内容精确率，以$(y \rightarrow y^i)$组合作为 BART 的输入和输出，得到的值反映了生成文本对参考文本的内容覆盖度。

5.2 无参考文本的评估方法

5.2.1 基于词序列统计特征的评估

该类评估方法利用文本自身的统计特征评估文本，如通过文本包含的词汇差异性反映文本多样性，利用词汇的组合概率反映文本的语言流畅性等。

Distinct-n[54]指标为生成文本中差异化词的数量和文本总的词汇数量的比值，值越大代表文本多样性越大：

$$Distinct-n(y)=\frac{|G_n(y)|}{\sum_{g\in G_n(y)}C(g,y)} \qquad (37)$$

Self-BLEU[53]指标通过多个生成文本间用词相似程度衡量生成文本多样性。针对一个输入文本所构成的生成文本集合Y，某个生成文本y的Self -BLEU 值为以该文本为目标文本，以集合中其他文本为参考文本计算得到的 BLEU 均值：

$$Self-BLEU(y)=\sum_{\bar{y}\in Y/y}BLEU(y,\bar{y})/|Y| \qquad (38)$$

Perplexity(PPL)[120]指标最初用于对比两个模型的性能，给高质量的文本赋予高概率值的模型较好。为了评估文本，通常对其变形使用，如下式：

$$PPL(y)=\sqrt[r]{\frac{1}{p(y)}} \qquad (39)$$

其中，r为生成文本长度，$p(y)$是文本y的出现概率，常用的计算方法是 N 元组的出现概率累乘：$p(y)=p(y_1)p(y_2|y_1)\cdots p(y_r|y_{r-N},\cdots,y_{r-1})$，每个 N 元组的概率通过在语料库中统计得到。PPL 值越低代表句子的出现概率越大，越符合人类语言习惯。

5.2.2 神经评估方法

该类方法借助神经网络从大规模的语料中学习语言学知识完成文本评估任务，如利用预训练语言模型中的语言知识评估文本的流畅度，利用预训练的 QA 模型的问答知识评估文本的内容正确性、或者利用推理模型评估给定两个文本的语义关联性。具体方法包括：

Forward Perplexity(FwPPL)[121]通过预训练的神经语言模型计算文本的出现概率，进而得到 PPL 值。神经语言模型的引入使 FwPPL 能够兼容统计语料中不常见句子对结果的影响。FwPPL 能反映单个文本质量，但如果所有的生成文本都和训练语言模型使用的语料中的某一个文本相似，那对于整个生成文本集合来说，FwPPL 会很低，表明生成文本质量很好，这和实际情况不符。

Reverse Perplexity(RevPPL)[121]针对 FwPPL 存在的问题，通过生成文本训练一个反向神经语言模型，然后由该语言模型计算正常句子的 PPL 值，只有具备多样性的生成文本所训练得到的语言模型，才能使正常文本的 PPL 值低，才能符合常识。该指标通常和 FwPPL 一起使用，从而能同时兼顾文本质量和文本多样性。

QAGS(Question Answering for Evaluation of Summaries)[122]从生成文本中选择多个实体和名词短语生成一个问题集合，使用预训练的问答模型分别以生成文本和输入文本为上下文回答问题，通过对比答案的一致性反映生成文本相比输入文本的事实一致程度。

roberta-sts[123]方法考虑具有语义关联的文本具有相互蕴含的关系，利用推理任务数据集 MNLI[124]中带有蕴含标签的数据训练roberta[125]模型，使其能够度量两个文本的语义关联，两个文本越倾向于蕴含关系，语义越接近。

ADEM 方法(Automatic Dialogue Evaluation Model)[126]使用人工生成的文本和模型生成的文本训练一个分类器，使其学习人类表达和机器表达的差异。由此，分类器判断对话回复是人写的还是机器生成。

通常情况下，人工生成的样本质量高于模型生成的样本质量，相同模型在同一个训练阶段生成的样本质量相似，但随着模型逐步被优化，训练后期生成的样本质量高于训练前期的样本质量。基于上述发现，CompEval(Comparator Evaluator)[127]方法标注不同生成源和不同训练阶段的样本，形成对比学习样本对，训练BERT区分两个文本的质量高低。

CTRLEVAL[128]评估方法构造属性相关的填充型提示模板，依据预训练语言模型对模板的填充结果，映射到对应属性，如为了判定一个句子的情感，构造如下模板：'y，it is [M]'，其中，y为生成文本，[M]为一个文本填充空间，预训练模型依据y的语义，在[M]处填充词汇，如果该词汇为积极情感，那么就判定y的情感极性为积极。

综上所述，评估方法经历了从具体词序列到抽象语义的发展过程，不同方法在评估粒度、核心思想等方面存在差异。我们汇总对比这些评估方法，如表4所示。

5.3 数据集

表5汇总了具备鲜明的类别标签如主题、情感极性、用户的属性等的数据集。现有的可控文本生成数据集类别较为单一，仍然缺乏一些能够反映自然语言细腻的情感表达和丰富的语言形式的标注样本，有待进一步构建具有细粒度标签的数据。

6 未来研究方向

可控文本生成是文本生成领域的重要研究方向，近年来相关理论和技术取得了长足进展。由于语言固有的灵活性、语义歧义等本质特性，以及缺乏多样性监督数据等现实约束，生成结果仍然存在可控属性单一、句子连贯性差、内容缺少个性化、违背常识等问题[151][152]。本章结合认知理论和计算技术探讨这些挑战和未来主要研究方向。

6.1 融合知识引导和数据驱动的可控文本生成

在可控文本生成任务中，由于监督数据难以覆盖所有符合给定约束的文本形式，导致监督学习方法缺少泛化能力。知识透过大量外在表象揭示事物间的本质关联，是人类可理解的，而且可以通过推理或者关联泛化到未见事物。因此，有必要探索融合数据驱动和知识引导的可控文本生成理论和技术，从而减小对标注数据的依赖和提升模型泛化能力。按照知识的表现形式，可以划分为结构化知识、文本型知识等，不同类型的知识在可控文本生成任务中有不同的应用方式。从认知科学角度，以知识图谱为代表的结构化知识是一种带有明确关联关系的知识形式，通常为包含实体和实体关系的复杂异构网络形式，如概念知识图谱、常识知识图谱、多语言词典知识库、领域知识图谱等等。为了利用结构化知识指导文本生成模型，首先需要依据输入文本及属性在知识图谱中检索到相关知识，然后将检索到的知识引入文本生成过程。检索结构化知识常用的方式包括语义检索、关键词匹配、属性匹配等，这些方法只能找到和输入文本直接关联的知识，难以发现并组织和输入文本存在深度语义关联的知识。因此，如何在特定语义和属性约束下，发现知识间的深度关联是一个重要方向。另外，为了将结构化知识引入文本生成过程，常用的方法是利用嵌入技术或者图注意力机制将图的语义信息编码入生成模型，但是图编码空间和文本编码空间是存在差异的。如何能在较少监督数据的条件下，缓解编码空间的差异，也是一个非常重要的研究方向。

相比结构化知识，文本型知识是一种更为常用和普遍的知识存在形式。预训练模型作为挖掘这些知识的主要手段，如何有效的利用预训练模型中和属性相关的知识非常重要。目前，常用的基于提示学习的可控文本生成方法通过设计和属性相关的前缀，使预训练模型自然地“回忆”起和属性相关的知识，但是仍然存在前缀的设计过度依赖专家知识、训练过程不稳定、难以找到最优解等缺点。如何有效的构建提示前缀、如何用和属性相关的前缀精细化地指导生成过程、针对提示前缀的训练策略等都是未来研究方向。另外，相比提示学习方式，是否有其他的更高效的方式能够激发预训练模型中的知识，仍然值得继续探索。

6.2 面向语言特征的文本评估

常用的文本评估方法或借助参考文本或单独训练一个神经网络评估模型。这些评估方法虽然能够大体反映文本质量，但是，在现实应用场景中，它们仍然面临非常大的挑战，主要体现在难以从语言学特征如拟人、反讽、比喻等修辞方法角度评估文本，距离具备“信、达、雅”的评估能力仍有很大的距离。

考虑到现有任务的评估需求很多涉及到非常抽象的语言学知识。因此，将语言学中对语言形式、语言含义和语境的分析成果和神经方法交叉融合是一个正确的大方向[153]。以往的语言学更倾向于对语言现象的解释，使得语言学和神经方法的融合较为困难，但随着语言学领域日渐对句法、语义和语用等特征的清晰描述和对这些特征间的使用规则的进一步总结，利用神经网络学习这些特征及其规则，进而评估文本成为可能。

表4 评估指标对比

评估分类	技术核心	评估指标	使用条件	评估粒度	核心思想
有参考文本的评估	元组匹配	BLEU[100]	无	词汇序列	生成文本元组相比参考文本元组的精确率
		NIST[101]	无	词汇序列	通过元组在参考文本中出现频次差异化元组权重
		ROUGE-n[102]	无	词汇序列	生成文本对参考文本的元组覆盖程度
		ROUGE-L[102]	无	词汇序列	考虑元组相对位置，引入子序列的匹配
		METEOR[103]	无	词汇序列	考虑元组相对位置，位置相邻的匹配元组更为重要
	编辑距离	TER[104]	无	词汇序列	词级别的编辑距离
		CharacTER[105]	无	字符序列	使用字符级别的编辑距离，减小同形词的影响
		ITER[106]	无	词汇序列	差异化编辑操作的代价
	词序列语义	EmbedAvg[107]	预训练的词向量	整体语义	在隐式空间表达句子语义，以词向量均值作为句子向量
		VecExt[108]		整体语义	在隐式空间表达句子语义，以词向量极值作为句子向量
		BERTScore[109]		词向量序列	基于融合上下文语义的词向量计算生成文本的精确率、召回率和 F 值
		WMD[111]		词汇序列	考虑两个文本中词向量的语义差异和词频差异
		WRD[112]		词汇序列	考虑两个文本中词向量的语义差异和重要程度差异
		SMS[113]		句子序列	以句子为最小评估单元，减小评估长文本的计算量
	神经评估	BLEURT[118]	构造训练数据	整体语义	拟合自动化评估指标和人工评估结果
		APES[116]	预训练 QA 模型	整体语义	基于 QA 模型判别两个文本的语义一致性
		QAEval[117]		整体语义	
		BARTScore[119]	预训练 BART 模型	整体语义	利用预训练模型挖掘文本间的语义关联
无参考文本的评估	统计特征	Self-BLEU[53]	同一个输入有多个生成文本	词汇序列	多个生成文本间互为参考文本
		Distinct-n[54]	统计所有词的词频	词汇序列	词汇差异越大，多样性越高
		PPL[120]	语料库	语言表达	句子出现概率越高越符合常规表达，基于大规模语料统计句子概率
	神经评估	FwPPL[121]	利用语料库训练语言模型	语言表达	句子出现概率越高越符合常规表达，基于神经语言模型计算句子概率
		RevPPL[121]	利用生成文本训练语言模型	语言表达	
		QAGS[122]	预训练 QA 模型	整体语义	基于 QA 模型的结果判定两个文本的语义一致性
		roberta-sts[123]	预训练推理模型	整体语义	语义一致的文本存在蕴含关系
		ADEM[126]	语料库	语言表达	训练二分类判别器学习机器生成文本和人工生成文本的表达差异
		CompEval[127]	构造正负样本	整体语义	通过对比学习使模型感知两个文本质量差异
		CTRLEVAL[128]	构造提示模板	整体语义	引入预训练模型中和属性相关的知识

表 5 可控文本生成数据集

任务	数据集	数据集来源	语言	数据集描述
主题、观点可控的文本生成	AGNews[1]	新闻文章	英文	100 万新闻文章，包含 4 类标签
	DBpedia[129]	新闻文章	英文	常用版包括 56 万训练数据和 7 万测试数据，14 类标签
	Switchboard-1 Release 2(SW2)[130]	通话记录	英文	70 个主题的 2400 条通话记录
	Switchboard Dialogue Act Corpus (SwDA)[130]	通话记录	英文	扩展了 SW2 的标签，标签包括陈述但无意见、陈述且有意见、同意、拒绝
	Yahoo Answers(YA)[131]	问答文本	英文	10 个主题的 448.3 万文本，每个文本包括问题、上下文、回答
	WikiSum[132]	维基百科	英文	233.2 万篇维基百科文章（视为摘要）及对应的引用文章
	Arxiv Academic Paper Dataset(AAPD)[2]	学术论文	英文	5.58 万篇学术论文摘要和主题，摘要平均长度 220 个词汇
	(New York Times)NYT[133]	新闻文章	英文	180 万新闻文章，其中超 150 万篇文章至少有一个标签，如主题、观点
	Chinese Poem Corpus (CPC)[134]	中文诗	中文	28.48 万古诗，其中 7.8 万篇 4 行诗
	ZHIHU corpus[135]	知乎问答	中文	5.5 万条主题词及对应的短文，短文长度在 50-100 个词之间
情感可控的文本生成	Internet Movie Database reviews(IMDB reviews)[136]	电影评论	英文	5 万带有正负情感标注的评论文本
	Stanford Sentiment Treebank (SST)[137]	电影评论	英文	SST-1 为 5 个情感类别的 1.18 万评论文本；SST-2 为 2 个情感类别的 0.9 万评论文本
	Beer Reviews (BR)[138]	啤酒评论	英文	158.6 万个评论文本、每个文本有外观、口感等的评分和整体评价
	Customer Reviews (CR)[139]	电子产品评论	英文	亚马逊网站上针对 5 类电子产品的带有正负情感标注的评论文本
	Daily Dialog(DD)[140]	对话文本	英文	1.3 万对话、平均长度 114.7 个词、并且人工标注了意图和情感
	Amazon Review (AR)[138]	产品评论	英文	AR-2 包括 2 个类别、400 万评论文本；AR-5 包括 5 个类别、70 万评论文本
	Yelp Reviews(YR)[131]	评论	英文	156.92 万条评论，每条评论包括评语和打分
个性化可控的文本生成	Twitter Dialogue Corpus (TDC)[141]	对话文本	英文	95 万对话、每个对话平均 6.27 个句子
	Ubuntu Dialogue Corpus (UDC)[142]	对话文本	英文	来自 Ubuntu 社区的 93 万条对话
	Cornell Movie Dialog Corpus (CMDC)[143]	电影字幕	英文	10292 个角色之间的 22 万对话、30.47 万个句子，带有角色属性
	OpenSubtitles dataset (OpS)[144] / SubTle Corpus[145]	电影字幕	英文	3600 万对话、1.4 亿句子/335 万对话、670 万句子
	PersonalDialog(PD)[146]	微博	中文	带有用户的特征（性别、年龄、位置、兴趣等）的 2000 万轮对话文本
问题类型、题材、句子角色等可控的文本生成	BookCorpus(BC)[147]	书籍	英文	16 个不同题材的 1.10 万书籍，共 7400 万个句子
	PubMed[148]	医学文摘	英文	20 万个论文-摘要对， 摘要每个句子标注了在摘要中角色，包括背景、目标、方法、结果和结论
	CNewSum[149]	新闻文章	中文	30 万个文档-摘要对，标注了文档中是否包含摘要的必要信息，也标注了摘要的信息能够从文档中推断出的难度
	DuReader[150]	百度搜索	中文	20 万不同类型的问题，每个问题对应多个回答，问题包括实体类，描述类，是非类三种类型，答案包括事实和观点类两种类型

1 http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html
2 https://github.com/lancopku/SGM

6.3 生成文本的可控性评价

由于自然语言的多样性，机械化的基于参考文本的评估指标通常难以准确地评估生成文本和属性的一致性，目前，基于判别器的方法和人工评估是可控文本生成领域的主要评估手段。但是，基于判别器的方法通常依赖于属性相关的数据集，而人工评估的主观性较大。未来研究可以关注于上述两个方面，一是设计准确可靠的面向文本可控性的定量评估指标，比如基于提示学习的属性评估方法。二是改进人工评估，提升评估的可靠性。

6.4 多方面细粒度的可控文本生成

传统的可控文本生成方法距离实际需求还有一定差距，主要体现在以下几个方面，一是属性粒度较大，不够精细。如在情感可控的文本生成任务中，通常只针对“积极”、“消极”等二元情感，这和人类多元且互相交叉的情感有很大的差异。二是属性单一，已有方法关注于某个特定属性，无法同时满足多个属性的控制。三是属性格式固定，不够灵活，大多数的模型使用一些简单变量，如使用 0、1 分别表达“科技”、“体育”等主题，但是，在现实场景中，人们往往倾向于用自然语言来表达更复杂化的控制需求。四是缺少和用户的交互，在生成过程中，属性对文本的约束无法动态调节。

因此，未来的研究方向可以关注于更细粒度且多方面的属性控制方法，具体包括更细腻的情感或者风格控制、生成同时满足多个属性约束的文本、以自然语言为控制变量的文本生成、在文本生成过程中引入可调参数动态控制文本属性等等。

6.5 文本幻觉问题

自然语言处理领域的幻觉问题是指文本内容存在不忠实于输入文本或者有违常识、存在虚假信息、难以验证真伪等情况。现有的基于神经网络的模型需要将输入文本映射到向量空间，借助在训练阶段学习到的文本语义相关性进行潜在推理，通过采样策略由向量空间映射回词汇空间。这一过程不同于人类的逻辑推理方式，缺少严格的逻辑判断，导致模型在语义空间难以全面理解输入文本的意图，生成与输入文本无关或者对立的幻觉文本。

为了减少文本幻觉，需要探索一些启发式方法，通过借鉴人类推理模式，辅助模型理解文本意图。概念和思维链是人类组织知识的核心。一方面可以借助输入文本中的概念和概念之间的关系，引入多任务学习[154]、结构化知识[155]等方式提升模型对输入文本语义的理解。另一方面，思维链(Chain-of-Thought，CoT)[156][157]是一系列中间推理步骤结果，常用于指导模型进行复杂推理。受此启发，可以将输入文本切分为模型可理解的形式，反映逻辑思维过程，从而降低模型理解人类复杂语言的难度，减少幻觉文本生成。

另外，在推理过程，可以探索如何引入常识知识来约束文本生成过程。现有的方法[122][154]通常关注生成文本和输入文本的事实一致性，缺少对生成文本和常识一致性的研究。基于常识知识的评估指标是一个重要方向，这类评估指标一方面可以在词汇采样阶段优化解码策略，避免选择有违常识的词汇，也可和基于思维链的方法相结合，度量模型推理过程的正确性，提升模型推理的可解释性。

6.6 文本偏见问题

偏见问题是指模型生成了不公平、歧视性等内容的文本，这会严重影响到人类对于生成系统的信任和使用。模型生成带有偏见文本的原因是使用了带有偏见的训练语料。因此，可以通过数据工程如转换训练数据中的性别等消除训练语料中的偏见。数据工程的方法依赖于专家知识，人工和再训练开销较大。也可从模型训练角度引入反映偏见程度的损失函数，如计算当前生成词汇和偏见词汇的距离，但会较大程度地牺牲语言流畅性。

控制偏见的前提是能够量化偏见，因此，未来研究方向可关注于如何设计公平合理的偏见量化指标。相比构造带有偏见程度标注的数据集训练神经网络，找到无偏见的文本更为容易。因此，可尝试的一种量化生成文本偏见的方法是对比其和无偏见文本的差异，这种差异可以从用词倾向性、情感等等不同视角度量。另外，偏见的程度往往依赖于文本主题等信息，一个文本在不同的主题下，其偏见程度也不尽相同，量化偏见时融入文本主题信息，从而在不同主题下差异化偏见程度，是更深层次的需要。

7 总结

可控文本生成是文本生成中的重要研究领域，因涉及复杂的语言学知识和认知本质极富挑战性，因其作为重要的人机交互形式极具实用价值。本文针对可控文本生成问题，综述代表性技术架构和模型，讨论这些框架和模型的基础理论和技术细节，针对文本多样性、长文本的语义一致性等高层次的控制需求，讨论相关技术的前沿进展，分析技术优势和不足。最后，讨论文本定性和定量评估指标，汇总可控文本生成的相关数据集，从认知和技术的角度探讨仍然存在的挑战和未来研究方向。

8 参考文献:

[1] Liu P, Yuan W, Fu J, et al. Pre-train, Prompt, and Predict: A systematic survey of prompting methods in natural language processing[J/OL]. arXiv preprint arXiv: 2107.13586, 2021.
[2] Wang K, Wan X. SentiGAN: Generating sentimental texts via mixture adversarial networks[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018: 4446-4452.
[3] 黄炎，孙海丽等. 基于主题约束的篇章级文本生成方法.北京大学学报, 2020, 56(1):9-15.
[4] 冯骁骋，龚恒等. 基于抽取的高考作文生成.计算机学报, 2020, 43(2):315-325.
[5] 杨锦锋,梁先桂,王刘安. 基于Prompt策略的医疗对话生成[J]. 中文信息学报. 2023, 37(4): 118-125.
[6] Kishinami Y, Akama R, Sato S, et al. Target-guided open-domain conversation planning[C]//Proceedings of the 29th International Conference on Computational Linguistics, 2022: 660-668.
[7] Tang Jian, Zhao T, Xiong C, et al. Target-guided open-domain conversation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 5624-5634.
[8] Madaan A, Setlur A, Parekh T, et al. Politeness transfer: a tag and generate approach[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 1869-1881.
[9] Prabhumoye S, Chandu K, Salakhutdinov R, et al. "My Way of Telling a Story": Persona based grounded story generation[J/OL]. arXiv preprint arXiv:1906.06401, 2019.
[10] Li D, Yang N, Wang W, et al. Unified language model pre-training for natural language understanding and generation[C]//Proceedings of the 33rd Conference on Neural Information Processing Systems, 2019:13063-13075.
[11] Thoppilan R, Freitas D, Hall J, et.al. Lamda: language models for dialog applications[J/OL]. arXiv preprint arXiv:2201.08239, 2022.
[12] Du Z，Qian Y, Liu X, et.al. GLM: General language model pretraining with autoregressive blank infilling[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022:320-335.
[13] Vaswani A, Shazeer N, et al. Attention is all you need[C]//Proceedings of the 31st Conference on Neural Information Processing Systems. Cambridge, 2017: 5998-6008.
[14] Webb, T, Holyoak, K, Lu, H. Emergent analogical reasoning in large language models[J/OL]. arXiv preprint arXiv:2212.09196, 2022.
[15] Bang Y, Cahyawijaya s, Lee N, et.al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity[J/OL]. arXiv preprint arXiv:2302.04023, 2023.
[16] Brown, T, Mann B, Ryder N, et al. Language models are few-shot learners[C]//Proceedings of the 34th Conference on Neural Information Processing Systems, 2020:7087-7097.
[17] Lewis M, Liu Y, Goyal N, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020:7871-7880.
[18] Radford, A, Wu, J, Child R, et.al. Language models are unsupervised multitask learners[R], USA: OpenAI, 2019.
[19] Open AI. GPT-4 Technical Report[R], USA: OpenAI, 2023.
[20] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research, 2020, 21:1-67.
[21] Ouyang L, Wu J, Jiang X, et.al. Training language models to follow instructions with human feedback//Proceedings of the 36th Conference Neural Information Processing Systems, 2022:27730-27744.
[22] Stiennon N, Ouyang L, Wu J, et.al. Learning to summarize from human feedback[J/OL]. arXiv preprint arXiv:2009.01325, 2020.
[23] Christiano P, Leike J，Brown T, et.al. Deep reinforcement learning from human preferences models[C]//Proceedings of the 31st Conference on Neural Information processing System, 2017:4299-4307.
[24] Mnih V, Badia A, Mirza M, et.al. Asynchronous methods for deep reinforcement learning[C]//Proceedings of the 33nd International Conference on Machine Learning, 2016:1928-1937.
[25] Schulman J, Levine S, Abbeel P, et.al. Trust region policy optimization[C]// //Proceedings of the 32nd International Conference on Machine Learning, 2016:1889-1897.
[26] Schulman, J, Wolski, F, Dhariwal, P, et.al. Proximal policy optimization algorithms[J/OL]. arXiv preprint arXiv:1707.06347, 2017.
[27] Dathathri S, Madotto A, Lan J, et al. Plug and play language models: a simple approach to controlled text generation[C]//Proceedings of the 8th International Conference on Learning Representations, 2020.
[28] Chan A, Ong Y, Pung B, et al. CoCon: A Self-supervised approach for controlled text generation[C]//Proceedings of the 9th International Conference on Learning Representations, 2021.
[29] Wei J，Bosma M，Zhao V, et.al. Finetuned language models

are zero-shot learners[C]// Proceedings of the 10th International Conference on Learning Representations, 2022.
[30] Sanh V, Webson A, Raffel C, et.al. Multitask prompted training enables zero-shot task generalization[C]// Proceedings of the 10th International Conference on Learning Representations, 2022.
[31] Zhou W, Jiang Y, Wilcox E, et. al. Controlled text generation with natural language instructions[C]//Proceedings of the 40th International Conference on Machine Learning, 2023.
[32] Zou X, Yin D, Zhong Q, et al. Controllable generation from pre-trained language models via inverse prompting[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021:2450-2460.
[33] Qian J, Li D, Yelong S, et al. Controllable natural language generation with contrastive prefixes[C]//Proceedings of the Findings of the Association for Computational Linguistics. 2022: 2912:2924.
[34] Li X, Liang P. Prefix-Tuning: optimizing continuous prompts for generation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021:4582-4597.
[35] Lester B, AI-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 3045-3059.
[36] Gu Y, Feng X, Ma S, et al. A distributional lens for multi-aspect controllable text generation[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022: 1023-1043.
[37] Zhang H, Song D. DisCup: Discriminator cooperative unlikelihood prompt-tuning for controllable text generation[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022: 3392-3406.
[38] Keskar N, McCANN B, Varshney L, et al. Ctrl: a conditional transformer language model for controllable generation[J/OL]. arXiv preprint arXiv:1909.05858, 2019.
[39] Yang K, Klein D. FUDGE: Controlled text generation with future discriminators[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021: 3511-3535.
[40] Pascual D, Egressy B, Meister C, et al. A plug-and-play method for controlled text generation[C]//Proceedings of the 2021 Conference Empirical Methods in Natural Language Processing, 2021:3973-3997.
[41] Gu Y, Feng X, Ma S, et al. Improving controllable text generation with position-aware weighted decoding[C]//Proceedings of the Findings of the Association for Computational Linguistics, 2022: 3449-4467.
[42] Goodfellow I, Pougetabadie J, et al. Generative adversarial nets[C]//Proceedings of the 28th Conference Neural Information Processing Systems, 2014: 2672-2680.
[43] Mirza M, Osindero S. Conditional generative adversarial nets[J/OL]. arXiv preprint arXiv: 1411.1784, 2014.
[44] Chen J, Wu Y, Jia C, et al. Customizable text generation via conditional text generative adversarial network[J]. Neurocomputing, 2020, 416: 125-135.
[45] 胡宇，王舰，孙宇清. 一种基于参考规范的专业文本生成方法[J]. 中文信息学报, 2021, 37(3):152-163.
[46] Zhang Y, Zhe G, et al. Generating text via adversarial training[C]//Proceedings of the 30th Conference on Neural Information Processing Systems, 2016.
[47] Jang E, Gu S, et al. Categorical reparameterization with gumbel-softmax[C]//Proceedings of the 5th International Conference on Learning Representations, 2017.
[48] Kusner M, Hernandezlobato J. GANS for sequences of discrete elements with the gumbel-softmax distribution[J/OL]. arXiv preprint arXiv: 1611.04051, 2016.
[49] Li J, Monroe W, Shi T, et al. Adversarial Learning for neural dialogue generation[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 2157-2169.
[50] Tuan Y, Lee H. Improving conditional sequence generative adversarial networks by stepwise evaluation[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2019, 27(4):788-798.
[51] Yu L, Zhang W, Wang J, et al. SeqGAN: Sequence generative adversarial nets with policy gradient[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence, 2017: 2852-2858.
[52] Fedus W, Goodfellow I, et al. MaskGAN: Better text generation via filling in the _[C]//Proceedings of the 6th International Conference on Learning Representations, 2018.1-15.
[53] Zhu Y, Lu S, Zheng L, et al. Texygen: a benchmarking platform for text generation models[C]//Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018:1097-1100.
[54] Li J, Galley M, Brockett C, et al. A diversity-promoting objective function for neural conversation models[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, 2016: 110-119.
[55] Lu S, Zhu Y, Zhang W, et al. Neural text generation: past, present and beyond[J/OL]. arXiv preprint arXiv: 1803.07133, 2018.
[56] Holtzman A, Buys J, et al. The curious case of neural text degeneration[C]//Proceedings of the 8th International Conference on Learning Representations, 2020.
[57] Gimpel K, Batra D, et al. A systematic exploration of diversity in machine translation[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013: 1100-1111.
[58] Vijayakumar A, Cogswell M, et al. Diverse beam search: decoding diverse solutions from neural sequence models[J/OL]. arXiv preprint arXiv: 1610.02424, 2016.
[59] Shao L, Gouws S, et al. Generating high-quality and informative conversation responses with sequence-to-sequence models[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 2210-2219.
[60] Shao Y, Gouws S, Britz D, et.al. Generating high-quality and informative conversation responses with sequence-to-sequence models[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017:2210-2219.
[61] Holtzman A, Buys J, Du L, et.al. The curious case of neural text degeneration[C]//Proceedings of the 8th International Conference on Learning Representations, 2020.

[62] Welleck S, Kulikov I, et al. Neural text generation with unlikelihood training[C]//Proceedings of the 8th International Conference on Learning Representations, 2020.
[63] Su Y, Lan T, Wang Y, et al. A contrastive framework for neural text generation[C]//Proceedings of the 36th Annual Conference on Neural Information Processing Systems, 2022: 21548-21561.
[64] Xu J, Ren X, Liu J, et al. Diversity-promoting gan: a cross-entropy based generative adversarial network for diversified text generation[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 3940-3949.
[65] Kingma D, Welling M. Auto-encoding variational bayes[C]//Proceedings of 2nd International Conference on Learning Representations, 2014.
[66] Sohn K, Lee H, et al. Learning structured output representation using deep conditional generative models[C]//Proceedings of the 29th Conference on Neural Information Processing Systems, 2015:3483-3491.
[67] Bao Y, Zhou H, Huang S, et al. Generating sentences from disentangled syntactic and semantic spaces[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 6008-6019.
[68] Shu L, Papangelis A, Wang Y, et al. Controllable text generation with focused variation[C]//Proceedings of the Association for Computational Linguistics, 2020: 3805-3817.
[69] Zhang X, Yang Y, Yuan S, et al. Syntax-infused variational autoencoder for text generation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 2069-2078.
[70] Yi X, Li R, Yang C, et al. MixPoet: diverse poetry generation via learning controllable mixed latent space[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020, 2020:9450-9457.
[71] Bowman S, Vilnis L, Vinyals O, et al. Generating sentences from a continuous space[C]//Proceedings of the 20th Conference on Computational Natural Language Learning, 2016:10-21.
[72] Dieng A, Kim Y, et al. Avoiding latent variable collapse with generative skip models[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence, CA: AAAI, 2019: 2397-2405.
[73] Makhzani A, Shlens J, et.al. Adversarial autoencoders[J/OL]. arXiv preprint arXiv:1511.05644, 2015.
[74] Zhang Y, Wang Y, Zhang L, et al. Improve diverse text generation by self-labeling conditional variational auto encoder[C]//Proceedings of the 44th International Conference on Acoustics Speech and Signal Processing, 2019: 2767-2771.
[75] Song T, Sun J, Chen B, et al. Latent space expanded variational autoencoder for sentence generation[J]. IEEE Access, 2019: 144618-144627.
[76] Fu H, Li C, Liu X, et al. Cyclical annealing schedule: a simple approach to mitigating kl vanishing[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019:240-250.
[77] Xu J, Durrett G. Spherical latent spaces for stable variational autoencoders[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 4503-4513.
[78] Hinton G, Srivastava N, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.
[79] Yang Z, Hu Z, Salakhutdinov R, et al. Improved variational autoencoders for text modeling using dilated convolutions[C]//Proceedings of the 34th International Conference on Machine Learning, 2017: 3881-3890.
[80] Ho J, Jain A, Abbeel P, et.al. Denoising diffusion probabilistic models[C]//Proceedings of the 34th Conference on Neural Information Processing Systems, 2020:6840-6851.
[81] Li X, Thickstun J, Gulrajani I, et.al. Diffusion-lm improves controllable text generation[J/OL]. arXiv preprint arXiv:2205.14217, 2022.
[82] Lovelace J, Kishore V, Wan C, et.al. Latent diffusion for language generation[J/OL]. arXiv preprint arXiv:2212.09462, 2022.
[83] Gong S, Li M, Feng J, et.al. DIffuseq: sequence to sequence text generation with diffusion models[C]//Proceedings of the 11th International Conference on Learning Representations, 2023.
[84] Li J, Luong M, Jurafsky D, et al. A hierarchical neural autoencoder for paragraphs and documents[C]//Proceedings of the 24th International Joint Conference on Natural Language Processing, 2015: 1106-1115.
[85] Zhao K, Ding H, Ye K, et al. A latent variable model with hierarchical structure and GPT-2 for long text generation[C]//Proceedings of the 30th International Conference on Artificial Neural Networks, 2021:297-308.
[86] Hu Z, Chan H, Liu J, et al. PLANET: dynamic content planning in autoregressive transformers for long-form text generation[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022:2288-2305.
[87] Huang Y, Xu K, Yu X, et al. Discourse-level text generation method based on topical constraint[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2020, 56(1): 9-15.
[88] Tan B, Yang Z, Al-Shedivat M, et al. Progressive generation of long text with pretrained language models[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, 2021:4313-4324.
[89] Li Q, Li P, Wei B, et al. Event transition planning for open-ended text generation[C]//Proceedings of the Findings of the Association for Computational Linguistics. 2022: 3412-3426.
[90] Mu F, Li W. Enhancing text generation via multi-level knowledge aware reasoning[C]//Proceedings of the 31st International Joint Conference on Artificial Intelligence, 2022: 4310-4316.
[91] Guo J, Lu S, Cai H, et al. Long text generation via adversarial training with leaked information[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018: 5141-5148.
[92] Liang X, Hu Z, Zhang H, et al. Recurrent topic-transition gan for visual paragraph generation[C]//Proceedings of the 2017 International Conference on Computer Vision, 2017: 3382-3391.
[93] Yang P, Li Lei, Luo F, et al. Enhancing topic-to-essay generation with external commonsense knowledge[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 2002-2012.
[94] Li X, Taheri A, et al. Commonsense knowledge base completion[C]//Proceedings of the 54th Annual Meeting of the

Association for Computational Linguistics, 2016:1445-1455.
[95] Sap M, Bras R, et al. Atomic: an atlas of machine commonsense for if-then reasoning[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018:3027-3035.
[96] Guan J, Huang F, Huang M, et al. A knowledge-enhanced pretraining model for commonsense story generation[J]. Transactions of the Association for Computational Linguistics, 2020, 8(1): 93-108.
[97] Guan J, Mao X, Fan C, et al. Long text generation by modeling sentence-level and discourse-level coherence[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021:6379-6393.
[98] Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019:3982-3992.
[99] Hua X, Sreevatsa A, et al. DYPLOC: dynamic planning of content using mixed language models for text generation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021:6408-6423.
[100] Papineni K, Roukos S, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002: 311-318.
[101] Doddington G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics[C]//Proceedings of the 2nd International Conference on Human Language Technology Research, 2002: 138-145.
[102] Lin C. ROUGE: A package for automatic evaluation of summaries[C]//Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 2004: 74-81.
[103] Lavie A, Agarwal A. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments[C]//Proceedings of the 2nd workshop on statistical machine translation, 2007: 228-231.
[104] Snover M, Dorr B, et al. A study of translation edit rate with targeted human annotation[C]//Proceedings of the 7th Conference of the Association for Machine Translation of the Americas, 2006: 223–231.
[105] Wang W, Peter J, Rosendahl H, et al. CharacTER: translation edit rate on character level[C]//Proceedings of the 1st Conference on Machine Translation, 2016:505-510.
[106] Panja J, Naskar S, et al. ITER: Improving translation edit rate through optimizable edit costs[C]//Proceedings of the 3rd Conference on Machine Translation, 2018: 746-750.
[107] Wieting J, Bansal M, et al. Towards universal paraphrastic sentence embeddings[C]//Proceedings of the 4th International Conference on Learning Representations, 2016.
[108] Forgues G, Pineau J, et al. Bootstrapping dialog systems with word embeddings[C]//Proceedings of the 28th NIPS workshop on Modern Machine Learning and Natural Language Processing, 2014.
[109] Zhang T, Kishore V, Wu F, et al. BERTScore: evaluating text generation with bert[C]//Proceedings of the 8th International Conference on Learning Representations, 2020.
[110] Devlin J, Chang M, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019: 4171-4186.
[111] Kusner M, Sun Y, Kolkin N, et al. From word embeddings to document distances[C]//Proceedings of the 32nd International Conference on Machine Learning, 2015: 957-966.
[112] Yokoi S, Takahashi R, et al. Word Rotator's Distance: decomposing vectors gives better representations[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020:2944-2960.
[113] Clark E, Celikyilmaz A, et al. Sentence Mover's Similarity: automatic evaluation for multi-sentence texts[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 2748-2760.
[114] Liu C, Lowe R, Serban I, et al. How not to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016: 2122-2132.
[115] Novikova J, Dusek O, et al. Why we need new evaluation metrics for NLG[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 2241-2252.
[116] Eyal M, Baumel T, Elhadad M. Question answering as an automatic evaluation metric for news article summarization[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, 2019:3938-3948.
[117] Deutsch D, Bedrax-Weiss T, Roth D. Towards question-answering as an automatic metric for evaluating the content quality of a summary[J]. Transactions of the Association for Computational Linguistics, 2021, 9(1):774-789.
[118] Sellam T, Das D, et al. BLEURT: learning robust metrics for text generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020:7881-7892.
[119] Yuan W, Neubig G, Liu P. BARTScore: evaluating generated text as text generation[C]//Proceedings of the 34th Neural Information Processing Systems, 2021:27263-27277.
[120] Jelinek F, Mercer R. Bahl L, et al. Perplexity -- a measure of the difficulty of speech recognition tasks[J]. Journal of the Acoustical Society of America, 1977, 62(S1).
[121] Zhao J, Kim Y, Zhang K, et al. Adversarially regularized autoencoders[C]//Proceedings of the 35th International Conference on Machine Learning, 2018: 5902-5911.
[122] Wang A, Cho K, Lewis M. Asking and answering questions to evaluate the factual consistency of summaries[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020:5008-5020.
[123] Kane H, Kocyigit Y, Abdalla A, et al. Towards neural similarity evaluators[C]//Proceedings of the 33rd Conference on Neural Information Processing Systems. Cambridge, 2019.
[124] Wang A, Singh A, Michael J, et al. Glue: a multi-task benchmark and analysis platform for natural language understanding[C]//Proceedings of the 2018 EMNLP Workshop, 2018:353-355.

[125] Liu Y, Ott M, Goyal N, et al. Roberta: a robustly optimized BERT pretraining approach[J/OL]. arXiv preprint arXiv: 1907.11692. 2019.
[126] Lowe R, Noseworthy M, Serban I, et al. Towards an automatic turing test: learning to evaluate dialogue responses[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1116-1126.
[127] Zhou W, Xu K. Learning to compare for better training and evaluation of open domain natural language generation models[C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020: 9717-9724.
[128] Ke P, Zhou H, Lin Y, et al. CTRLEval: an unsupervised reference-free metric for evaluating controlled text generation[C] //Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 2306-2319.
[129] Jens L, Robert I, et al. DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia[J]. Semantic Web, 2015, 6(2): 167-195.
[130] Godfrey J, Holliman E. Switchboard-1 release 2[R], Philadelphia: Linguistic Data Consortium, 1993.
[131] Zhang X, Zhao J, et al. Character-level convolutional networks for text classification[C]//Proceedings of the 29th Conference on Neural Information Processing Systems, 2015: 649–657.
[132] Liu P, Saleh M, et al. Generating wikipedia by summarizing long sequences[C]//Proceedings of the 6th International Conference on Learning Representations, 2018.
[133] Sandhaus E. The New York Times Annotated Corpus LDC2008T19[R]. Philadelphia: University of Pennsylvania, 2008.
[134] Zhang X, Lapata M. Chinese poetry generation with recurrent neural networks[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 670-680.
[135] Feng X, Liu M, Liu J, et al. Topic-to-essay generation with neural networks[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018: 4078－4084.
[136] Diao Q, Qiu M, Wu C, et al, Jointly modeling aspects, ratings and sentiments for movie recommendation(JMARS)[C]//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2014:193–202.
[137] Socher P, Perelygin A, et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013:1631-1642.
[138] McAuley J, Leskovec J. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews[C]//Proceedings of the 22nd International Conference of World Wide Web, 2013:897–908.
[139] Hu M, Liu B, et al. Mining and summarizing customer reviews[C]//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2004:168–177.
[140] Li Y, Su H, Shen X, et al. Dailydialog: a manually labelled multi-turn dialogue dataset[C]//Proceedings of the 8th International Joint Conference on Natural Language Processing, 2017:986-995.
[141] Ritter A, Cherry C, et al. Data-driven response generation in social media[C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011:583–593.
[142] Lowe R, Pow N, et al. The ubuntu dialogue corpus: a large dataset for research in unstructured multi-turn dialogue systems[C]//Proceedings of the 16th Annual SIGDIAL Meeting on Discourse and Dialogue, 2015:285-294.
[143] Danescu-Niculescu-Mizil C, Lee L. Chameleons in imagined conversations: a new approach to understanding coordination of linguistic style in dialogs[C]//Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics, 2011:76-87.
[144] Lison P, Tiedemann J. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles[C]// Proceedings of the Tenth International Conference on Language Resources and Evaluation, 2016:923-929.
[145] Ameixa D, Coheur L. From subtitles to human interactions: introducing the subtle corpus[R]. Lisboa: INESC, 2013.
[146] Zheng Y, Chen G, Huang M, et al. Personalized dialogue generation with diversified Traits [J/OL]. arXiv preprint arXiv: 1907.11692. 2019.
[147] Zhu Y, Kiros R, et al. Aligning books and movies: towards story-like visual explanations by watching movies and reading books[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision, 2015:2380-7504.
[148] Dernoncourt F, Lee J. PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts[C]//Proceedings of the 8th International Joint Conference on Natural Language Processing, 2017:308-313.
[149] Wang D, Chen J, Wu X, et al. CNewSum: a large-scale chinese news summarization dataset with human-annotated adequacy and deducibility level[c]//proceedings of the 10th CCF International Conference on Natural Language Processing and Chinese Computing, 2021:389-400.
[150] He W, Liu K, Liu J, et al. DuReader: a Chinese machine reading comprehension dataset from real-world applications[C]//Proceedings of the Workshop on Machine Reading for Question Answering, 2018:37-46.
[151] Li H, Zhu J, Zhang J, et al. Ensure the correctness of the summary: incorporate entailment knowledge into abstractive sentence summarization[C]//Proceedings of the 27th International Conference on Computational Linguistics, 2018:1430-1441.
[152] 黄民烈，万小军，高扬. 中文信息处理发展报告[R]. 北京：中国中文信息学会, 2021.
[153] 陆俭明. 亟需解决好中文信息处理和汉语本体研究的接口问题[J]. 当代修辞学, 2021, (1):1-10.
[154] Chen X, Li M, Gao X, et.al. Towards improving faithfulness in abstractive summarization[C]//Proceedings of the 36th Conference on Neural Information processing System, 2022: 24516-24528.
[155] Wu Z, Galley M, Brockett C, et.al. A controllable model of grounded response generation[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2021: 14085-14093.
[156] Wei J, Wang X, Schuurmans D, et.al. Chain-of-Thought prompting elicits reasoning in large language models[C] //Proceedings of the 36th Conference on Neural Information Processing Systems, 2022:24824-24837.
[157] Kojima T, Gu S, Reid M, et.al. Large language models are

zero-shot reasoners[C]//Proceedings of the 36th Conference on Neural Information Processing Systems, 2022: 22199-22213.

王舰(1991－),博士生，主要研究方向为语义计算。
E-mail:
wangjian026@126.com

孙宇清(1967－),通讯作者,博士， 教授,中国计算机学会杰出会员,主要研究领域为语义与协同计算、自然语言处理。
E-mail:
sun_yuqing@sdu.edu.cn