基于内在质量约束的文本生成和评价综述

兰玉乾[1],饶 元[1*],李冠呈[2],孙 菱[1],夏昺灿[1],辛婷婷[1]
(1. 西安交通大学软件学院社会智能与复杂数据处理实验室,陕西西安 710049;2. 中国长峰机电技术研究设计院,北京 100854)

摘 要: 近年来,以ChatGPT为代表的能够适应复杂场景、并能满足人类的各种应用需求为目标的文本生成算法模型成为学术界与产业界共同关注的焦点. 然而,ChatGPT等大规模语言模型(Large Language Model,LLM)高度忠实于用户意图的优势隐含了部分的事实性错误,而且也需要依靠提示内容来控制细致的生成质量和领域适应性,因此,研究以内在质量约束为核心的文本生成方法仍具有重要意义. 本文在近年来关键的内容生成模型和技术对比研究的基础上,定义了基于内在质量约束的文本生成的基本形式,以及基于"信、达、雅"的6种质量特征;针对这6种质量特征,分析并总结了生成器模型的设计和相关算法;同时,围绕不同的内在质量特征总结了多种自动评价和人工评价指标与方法. 最后,本文对文本内在质量约束技术的未来研究方向进行了展望.

关键词: 自然语言处理;语言模型;文本生成;文本质量;文本评价

基金项目: 国家自然科学基金重点项目(No.U22B2036);科技部重点研发计划项目(No.2019YFB2102300);中央高校建设世界一流大学(学科)和特色发展引导专项资金项目(No.PY3A022)

中图分类号: TP183 **文献标识码:** A **文章编号:** 0372-2112(2024)02-0633-27

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230826

A Survey of Text Generation and Evaluation Based on Intrinsic Quality Constraints

LAN Yu-qian[1], RAO Yuan[1*], LI Guan-cheng[2], SUN Ling[1], XIA Bing-can[1], XIN Ting-ting[1]
(1. *Laboratory of Social Intelligence & Complex Data Processing, School of Software Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi* 710049, *China*;
2. *Beijing China Changfeng Electromechanical Technology Research and Design Institute, Beijing* 100854, *China*)

Abstract: Recently, the outstanding text generation language models represented by ChatGPT, which can adapt to complex scenes and meet various application demands of human beings, has become the focuses of both the academic and industrial circles. However, the advantage of large language models (LLM) such as ChatGPT that are highly faithful to user intent implies some factual errors, and it is also necessary to rely on prompt content to control the detailed generation quality and domain adaptability, so it is still of great significance to study text generation with intrinsic quality constraints as the core. Based on the comparative study of key content generation models and technologies in recent years, this paper defined the basic form of text generation with intrinsic quality constraints, and six quality features based on "credibility, expressiveness and elegance". In view of these 6 quality features, we provided analysis and comparison of generator model design and related algorithms. Besides, various automatic and human evaluation methods for different intrinsic quality features are summarized. Finally, this paper looks forward to the future research directions of intrinsic quality constraint technology.

Key words: natural language processing; language model; text generation; text quality; text evaluation

Foundation Item(s): National Natural Science Foundation of China (No. U22B2036); National Key Research and Development Program of China (No. 2019YFB2102300); Special Funds Program for Central Universities to Build World-class Universities (Disciplines) and Guide the Development of Special Features (No.PY3A022)

收稿日期:2023-09-01;修回日期:2024-01-02;责任编辑:孙瑶
*通讯作者:饶元

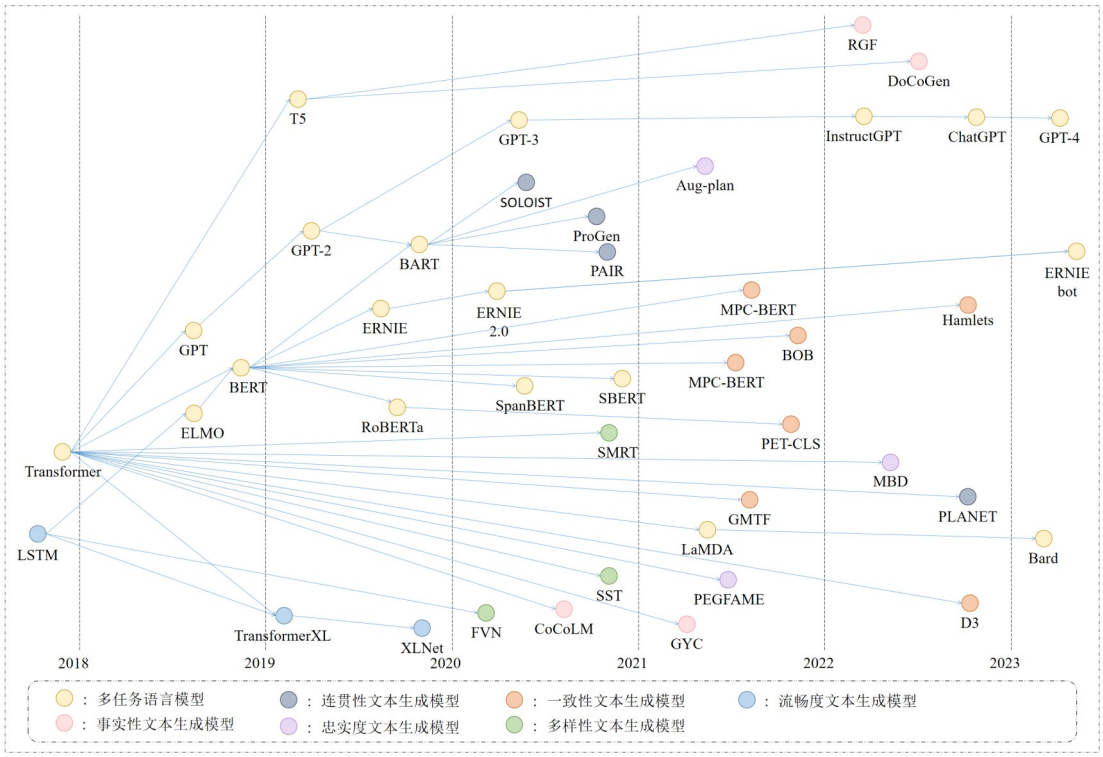
1 引言

2022年11月30日,由OPENAI实验室推出的一款基于GPT3.5的内容生成工具ChatGPT[1,2],通过进一步提升模型的记忆能力与文本理解能力,使其自动生成的问题解答、软件代码、数学计算和信件内容等结果,不仅具备优秀的内容完整性和逻辑性,而且能够符合用户偏好并实现场景的自适应性,从而引起了学术界和工业界的广泛关注,并使基于人工智能的内容生成技术(AI Generate Content,AIGC)成为目前AI技术领域中热议的焦点之一[3,4]. AIGC技术的核心是通过AI算法自动化地生成满足特定目标和质量要求的文本内容,生成文本不仅需要符合图灵假设[5],而且应当满足人们在浏览信息时所需要的"信(Credible)、达(Expressiveness)、雅(Elegance)"的质量需求[6]. 目前,以GPT-3[3],T5[7]和GPT-4[8]为核心的大规模预训练语言模型(Large Language Model,LLM)不仅具有的层数多、参数量大的结构特点,而且通过海量语料的训练,具有很强的理解能力和泛化能力,逐渐成为AIGC技术的主流. 但是,相比LLM聚焦不同外在环境和任务的普适性,如何提升语言模型所生成文本的内在质量以符合特定领域的质量需求,迫切需要开展深入的研究.

近年来语言生成模型的技术演化路线如图1所示. 其中,Google和OpenAI的研究者在早期基于多头注意力机制的Transformer[9]模型基础上,通过增加预训练子任务或改进解码器结构,分别提出了用于自然语言理解任务的BERT模型[10]和适应多种任务的自回归模型GPT[11]. 此后预训练语言模型出现了两个重要分支:一是以GPT-2[12],GPT-3[3],T5[7]和GPT-4[8]为代表的自回归模型采用了更加丰富的训练语料和更加庞大的参数,进一步增强了LLM在多种不同生成任务中的性能与泛化性;二是以RoBERTa[13],SpanBERT[14]和SBERT[15]为代表的改进模型通过改进掩码机制和预训练子任务,进一步提升了BERT模型的编码性能和应用领域. 同一时期,基于循环机制和双流自注意力的TransformerXL[16]和XLNet[17]等预训练模型在生成长文本的同时,通过生成质量控制保持了上下文语义的流畅性;而融合了BERT编码能力和GPT-2生成优势的BART模型[4]能够通过输入不同的关键字来控制生成文本的内容语义. 为了生成更加连贯的长文本内容,Tan等人[18]基于BART构建了多步骤生成模型ProGen,通过在多步生成过程中采用不同的触发词,有效控制了长文本生成中内容语义的连贯性,并在基于CNN新闻数据集的实验结果表明,相对于BART模型的BLEU值30.1%,ProGen模型提升到了31.2%. 而Hua等人[19]从语义连贯性与动态语义的演化特征出发,提出了连贯性生成模型PAIR,它将文本计划作为生成模型的输入以控制输出文本的整体脉络,从而保证了输出内容具有更加平滑的语义变化和连贯性,在针对Reddit数据集的测试中发现,相比BART模型的BLEU值6.78%,PAIR的BLEU值提升至36.09%. Jang等人[20]为了增强对话任务中回复内容的角色一致性,提出的模型将BART模型作为生成器的主体结构,采用角色和知识的独立编码以增强回复文本中的属性表达,该模型在FoCus数据集上测试结果为46.31%,比BART模型的BLEU值13.18%大幅提升. 上述工作表明,从不同的质量特征上来优化内容生成模型的结构,均可获得较BART模型更优的结果,因此,在各种大语言模型不断推新的场景下,从质量特征的视角来进行模型优化,仍然具有的重要的研究价值和意义.

围绕着生成更准确[21,22]、更真实[23,24]、更细致[25~27]、更可靠[28,29]的高质量文本内容,大量学者从不同的视角对AIGC领域中的关键技术问题进行了综述研究. 其中,Iqbal等人[30]在CNN,RNN,LSTM等经典模型的基础上,深入分析和对比了基于VAE和GAN等生成模型的差异,发现传统语言模型无法控制不同文本质量中的细微差别. Li等人[31]则从预训练语言模型和微调机制出发,从"相关性、忠实度、保序性"等质量维度来分析不同类型的输入数据如何自适应地生成满足特定质量要求的文本,但是这些工作缺少探索所选特性的可泛化能力,导致一些质量特征仅限于特定的任务. 在生成文本的可靠评价角度上,Celikyilmaz等人[32]从人工评价方法、非训练的自动评价算法和机器学习的评价模型等方向出发,系统讨论了生成文本的评价方法与指标体系. 而Jin等人[33]关注文本风格迁移任务,特别是围绕着风格迁移强度、语义保存性和流畅度等特征深入研究了自动评价和人工评价方法. 然而,现有的评价方法依然偏向于特定的生成任务,缺乏一个以质量特征为核心主线的内容评价框架.

综上所述,在考虑到不同研究工作之间关注焦点存在的差异性,本文在梳理和对比与已有研究综述在AIGC核心问题与工作挑战的基础上(表1),从质量约束与控制的视角出发,对AIGC中高质量文本内容生成进行形式化定义,进而围绕不同的质量特征与任务分析相关技术、模型的研究进展,对未来的发展趋势进行分析和总结,为未来的研究奠定基础和指引.



注:上述演化模型都关注不同质量特征的约束.

图 1　文本生成模型演化示意图

表 1　相关综述和研究工作的对比表

对比文献	主要工作	关键问题	相关章节
Iqbal 等人[30]	围绕各种经典深度学习和文本生成模型进行介绍和总结.	经典文本生成模型之间的差异是什么?	3
Li 等人[31]	介绍了预训练语言模型(Pre-trained Language Model, PLM)的主流架构,并分析相关模型设计,以及微调策略.	如何编码输入数据,设计和优化预训练语言模型作为生成器?	3
Zhou 等人[34]	提出释义生成的技术架构,并讨论生成流畅、多样化和拟人的释义的方法.	如何产生符合质量要求的释义内容?	3
Zhao 等人[35]	全面介绍了基于主题的文本生成技术.	如何构建基于主题的文本生成模型?	3
Ji 等人[36]	介绍了以幻觉为代表的忠实度和事实性错误,以及不同的缓解方法.	幻觉的定义、原因、缓解方法都是什么?	3
Tang 等人[37]	介绍了文本生成的数据构建、模型框架、训练和推断策略以及评估方法.	神经网络模型的相关改进有哪些?	3
Celikyilmaz 等人[32]	介绍了文本的人工评价方法、非训练的评价方法和自动评价模型.	如何评估文本生成?	4
Jin 等人[33]	讨论文本迁移任务中的定义、数据集、任务和评估方法.	如何定义和评价文本风格迁移?	3, 4
本文研究	① 给出了基于内在质量约束的生成定义,以及核心质量特征; ② 讨论了不同质量特性条件下的文本生成模型设计或算法; ③ 分析了不同质量特征对应的评价方法.	如何针对每种质量特征,总结相关的生成器模型设计或算法? 如何针对不同质量特征,分析相关的自动和人工评价方法?	2, 3, 4

2　问题与挑战

定义 1　基于内在质量约束的文本生成

$$p\left(Y\left|X,Q\right.\right)=\prod_{i=1}^{N}p\left(y_i\left|X,Q,y_1,y_2,\cdots,y_{i-1}\right.\right)\quad(1)$$

其中,$p\left(Y\left|X,Q\right.\right)$表示生成器$G$在输入$X$并融合特定内在质量特征$Q$,自动生成长度为$N$的文本$Y$的概率.$p\left(y_i\left|X,Q,y_1,y_2,\cdots,y_{i-1}\right.\right)$表示输入$X$在第$i$时刻受到$Q$以及已生成的历史序列$y_1,y_2,\cdots,y_{i-1}$的约束影响下,生

成预测词y_i的条件概率，而$\prod_{i=1}^{N} p\left(y_i \mid X, Q, y_1, y_2, \cdots, y_{i-1}\right)$表示基于$X$、$Q$和历史序列$y_1, y_2, \cdots, y_{i-1}$生成长度为$N$的词序列$y_1, y_2, \cdots, y_i$的联合概率. 式(1)表示基于文本内在质量约束的基本生成过程，不同质量特征Q的具体表达有不同的形式，包括关键词、证据、风格等. 针对不同质量特征Q的生成器G也需要适应性的设计调整，包括编码机制、解码器结构、生成步骤等. 基于文本翻译“信、达、雅”的质量要求，本文将其迁移并映射到式(1)的质量特征Q中，得到了6种基本形式，如表2所示.

表2中，“信”可以被理解为生成文本被特定“信息源”所支持且不存在矛盾，一旦所生成的文本不符合“信”的质量要求时，可能出现“幻觉”[36]. 根据“信息源”本身的可信程度又将“信”划分为忠实度和事实性：当生成模型仅关注“支持信息源”而不考虑“信息源”是否为客观事实时，“信”被表达为忠实度，此时生成文本需要符合或覆盖输入内容，且限制未被输入内容所涵盖或证实的幻觉信息；当“信息源”代表了客观事实或常识知识时，“信”则表示为事实性，此时生成文本需要与某个客观证据相匹配. 而“达”可以理解为实现上下文语义相互关联的内容组织，从动态和静态的角度出发又可以将“达”分为连贯性和一致性. 连贯性关注生成文本在某个范畴内的句子语义是否动态变化且连续[38]，而一致性则聚焦生成文本的上下文静态属性是否保持不变，如用户角色属性或内容情感属性等. “雅”是文本质量的高级要求，不同的研究者对“雅”的定义和关注点不同，本文将“雅”分为多样性和流畅度. 多样性表示基于相同语义可以具有多种不同文本表达，从而提升了不同任务下的适应性；而流畅度则关注于生成文本中前后语句搭配是否合理.

基于内在质量约束的文本生成和上述6种质量特征Q的定义，以及文本生成效果的评价方法，本文构建了一个基于内在质量约束的整体研究框架，并归纳出2个核心的研究问题，如图2所示.

表2 文本质量特征描述

质量要求	质量特征	描述	关键要素
信(Credibility)	忠实度(Faithfulness)	生成文本覆盖输入内容,避免出现未被输入内容所支撑的幻觉信息.	输入内容
	事实性(Factuality)	生成文本满足客观事实或常识逻辑.	证据
达(Expressiveness)	连贯性(Coherence)	生成文本满足特定范畴,范畴内的语义动态变化且连续.	范畴
	一致性(Consistency)	生成文本所具有的静态属性在有限的上下文中不发生突变.	属性
雅(Elegance)	多样性(Diversity)	生成文本在满足语义内容的同时具有不同的表达.	风格
	流畅度(Fluency)	生成文本中前后语句搭配合理且具有信息量.	N-gram

研究问题1：基于不同的内在文本质量特征Q的描述，如何设计生成器G的模型或相关算法？

研究问题2：对于不同质量特征Q的定义，如何构建针对性的文本质量评价框架？

虽然以ChatGPT[2]和CPT-4[8]为代表的通用人工智能(Artificial general intelligence，AGI)模型已经成为学术界和工业界的技术标杆，但以生成质量为导向的中小规模语言模型研究仍然具有较大的价值，原因有三个方面：从质量角度看，ChatGPT等模型仍然存在事实性和一致性不足；从任务角度看，确保超大语言模型符合各类任务的前提仍然是高质量的提示内容(prompt)；从领域角度看，得到特定数据集训练的中小规模语言模型也能够达到或超过ChatGPT等模型的生成质量，且易于微调.

3 文本内在质量约束研究现状分析

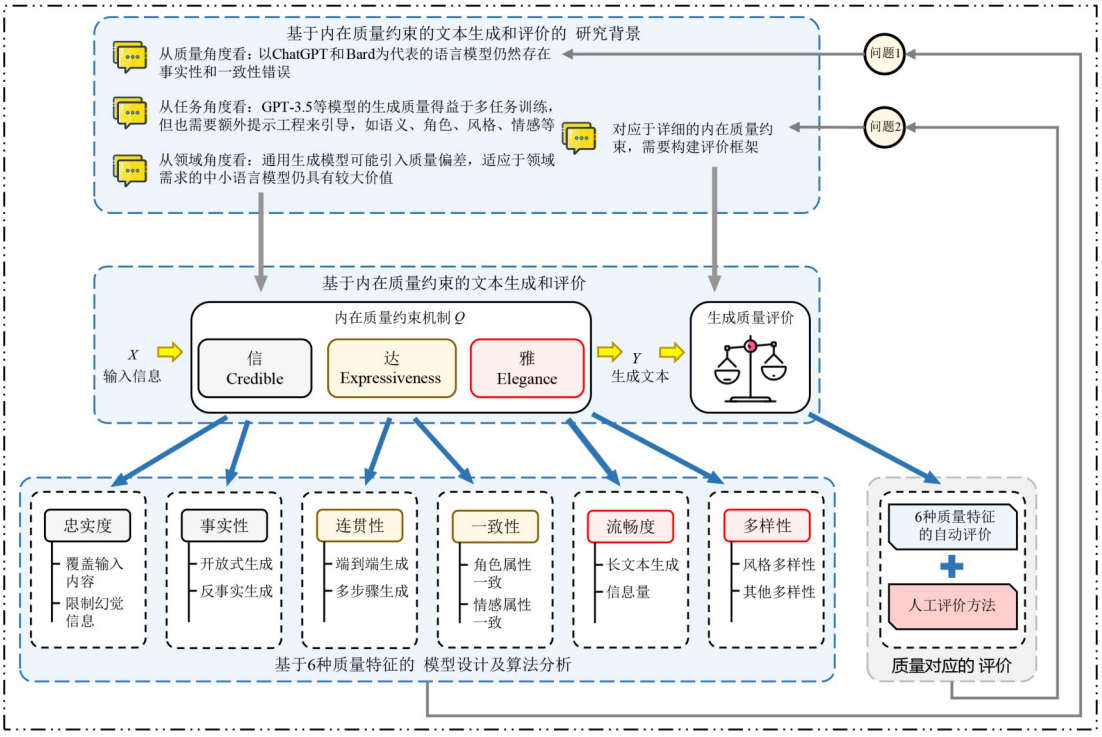
基于内在质量约束的文本生成定义以及6种质量特征Q的描述，本节将分别讨论相关模型的结构设计和技术原理，并做出分析比较. 考虑到基于词重叠的评价方法便于计算且被大部分生成模型使用，本节主要利用BLEU值、Accuracy值和F1值等指标对不同模型的生成质量进行对比.

3.1 忠实度

忠实度表示生成文本应当覆盖输入内容(如文档、句子、表格)等关键信息，同时限制生成内容不包含其他的幻觉信息. 如图3所示，基于忠实度的典型生成模型，忠实度的核心问题包括如下两个方面：一是如何覆盖所有“输入内容”的关键语义或实体；二是如何限制幻觉信息对生成文本的影响. 针对上述这两个问题，本节对基于忠实度的文本生成模型与技术的当前研究进展进行分析，相关模型如表3所示.

3.1.1 输入内容特征的覆盖度

覆盖输入内容特征的基本思想是将输入文本的内容特征编码为有组织的词序列，再解码为符合忠实度要求的文本. 训练数据中往往存在一定的噪声，会导致输入内容和生成文本之间的实体词无法对齐或内容不匹配，影响模型训练后的最终生成结果，因此，为了保证和提升对输入内容特征的覆盖率，相关研究的思路可以分为训练集的数据增强与输入信息的编码控制这两个部分.



注:从内在质量约束的研究背景出发,将“信,达,雅”映射为6种质量特征以及评价方法.

图2　基于内在质量特征约束的文本生成和评价框架

针对训练集数据增强的核心方法是构造训练样本的忠实度标签,并将忠实度标签和输入文本拼接成为新的训练样本. 例如,Rebuffel等人[39]提出了一个基于词级忠实度标签的MBD模型,该模型基于词共现和句法依赖树来计算词级忠实度的分值(图3(*b*)),通过词级忠实度标签来增强原始训练数据集,并指导分枝解码器按照不同的权重生成符合忠实度要求的单词(图3(*e*)),以提升生成文本用词的精确度;而Rashkin等人[40]将这种思想应用于句子级的训练样本中,即通过句子的“表达客观性”“覆盖率”“蕴含率”三种评价指标来分别构造训练样本的句子级忠实度标签(图3(*c*)),并将三种标签与输入文本拼接作为新的训练数据. 这两种方法在基于Wikipedia数据集进行数据增强实验的结果表明:相比句子级标签的数据增强方法[40],采用词级标签的数据增强方法[39]的BLEU值从之前的8.9%提升至了41.56%. 这表示明词级标签比句子级标签更加细致,对生成文本的忠实度的质量控制更好.

针对输入信息的编码控制则是通过编码器提取出输入文本中的关键信息,并重新组织生成核心的内容序列. 在文本摘要生成任务中,Li等人[41]采用共享编码器的多任务学习在文本生成任务的基础上增加输入信息的分类任务和反馈训练,以聚焦输入内容的关键语义. 而基于表格数据到文本的生成中,需要提取核心实体和属性. Liu等人[42]提出的Aug-plan模型将结构化的表格数据转化为文本序列(图3(*d*)),再通过伪平行数据将文本序列进一步扩大为文本计划(plan),该文本计划中包含了能够覆盖输入表格中的所有实体词和属性值信息,可直接控制生成文本的核心内容. 在基于WIKIPERSON数据集的测试中发现,相比朴素的BART模型[4],Aug-plan模型的PARENT值从52.54%提升到了56.75%. 但是,Aug-plan模型中使用的文本计划虽然能够有效覆盖表格数据,也可能会引入伪平行数据中的幻觉信息而产生噪声.

3.1.2　限制幻觉信息的生成

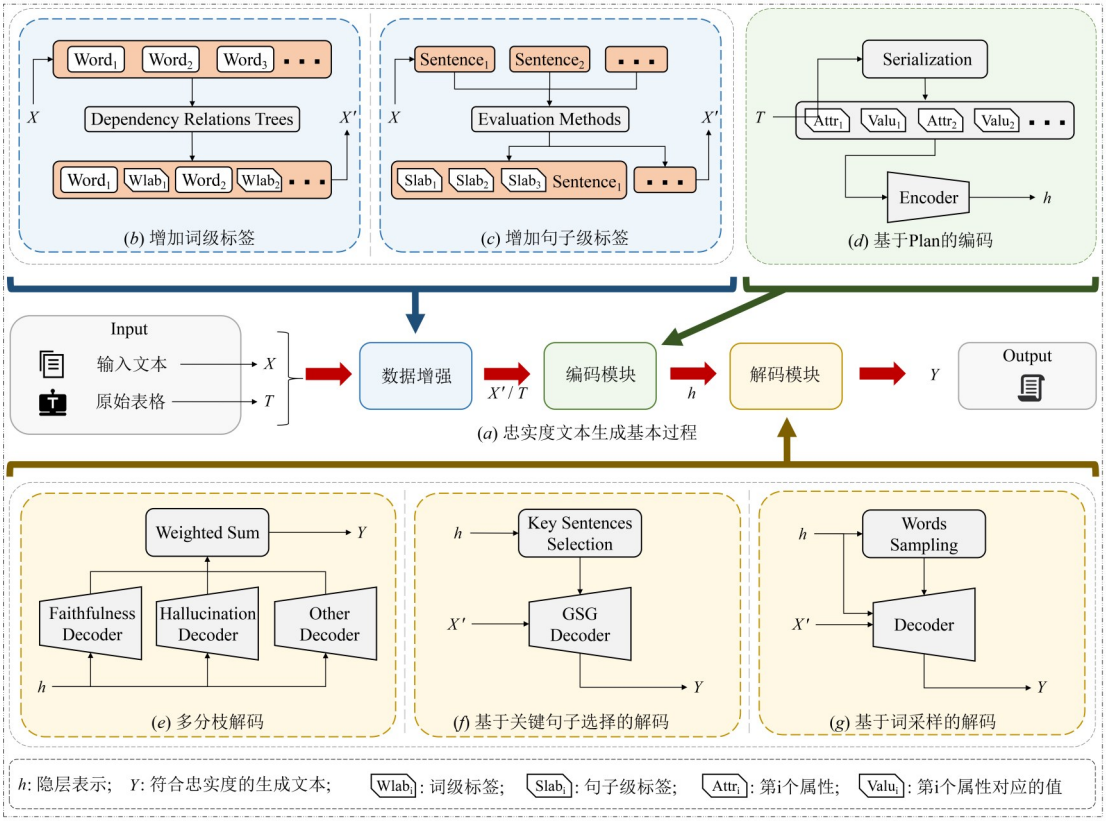
限制幻觉信息的核心任务是避免预训练数据中潜在的幻觉实体词对生成结果产生干扰. 为了保证生成摘要完全符合输入文档的要求,Zhang等人[43]提出了PEGASUS模型,通过采用间隔句子生成(Gap Sentences Generation,GSP)子任务来实现对输入文档中的部分句子进行掩蔽,并采用句子重要性选择模型来对关键句子进行预测,从而实现了抽取式摘要生成的目标(图3(*f*)). PEGASUS模型的本质是从输入文档中选择句子并构成摘要,这种方法能够很好地限制幻觉信息,但是句子之间的连贯性将会损失. 在此基础上,Aralikatte等人[44]进一步采用词级相似性和上下文注意力机制从输入内容中提取焦点词作为解码器的输入,并通过生成式摘要方法设计出了PEGFAME模型,实现了输入文本内容编码的严格控制(图3(*g*)). 在XSUM数据集上的实验

结果表明，PEGFAME 模型在 PEGASUS 模型的 BERT-Faithful 分值 27% 的基础上进一步提升了 0.3%. 同时，PEGASUS 模型和 PEGFAME 模型都将幻觉信息比率均控制在了 0.26%，即聚焦输入文本中的核心句子和核心词，以回避训练样本中可能存在的噪声信息，但这种焦点编码方法也会导致生成文本对输入内容的覆盖率不足.

此外，对比学习也被用于模型训练以避免幻觉信息的出现，Cao 等人[45]和 Lee 等人[46]构建了用于区分幻觉信息的正样本和负样本集合. 其中，正样本来源于同义词替换、随机单词替换和回译，而负样本则通过实体词更换、MASK 填充、增加低概率内容和构造对抗样本等方式引入幻觉. 基于正负样本的对比学习损失有助于模型提升忠实度.

3.1.3 小结

上述研究工作对应于忠实度的两项核心问题，并采用了完全不同的两种思想. 其中，改造数据集或增加标签的优点在于促进了模型对于关键语义的聚焦[39,40]，但依赖模型的编解码设计，如多任务共享编码器、数据序列化处理和分值解码器等，而且数据集样本的增强或标签的准确性都将限制整体模型的性能；采用实体词构成文本计划的方法能够保证覆盖大部分输入信息的同时兼顾生成内容的语义连贯性和内容逻辑[42]，但缺点是解码阶段中可能掺杂幻觉. 所以，在解码阶段前抽取关键句子和关键词以实现焦点编码能够进一步避免幻觉信息的引入[43,44]，但这种抽取机制也会损失一定的语义信息和连贯性. 通过对比学习能够改进模型对幻觉的辨别能力以提高忠实度[45,46]，但如何在保证原始样本语义不变的前提下，构造可靠的正样本以及包含多种幻觉类型的负样本也依赖独立的样本处理算法或细致的人工操作.



注：(*a*)表示此类模型的基本生成过程，(*b*)和(*c*)分别通过词级和句子级标签使模型关注于输入内容的关键词句，(*d*)将输入内容中的关键信息转化为文本计划，(*e*)通过分枝解码控制生成内容的忠实度，(*f*)和(*g*)通过关键句子抽取和关键词采样控制生成内容的忠实度.

图 3 基于忠实度的典型生成模型

3.2 事实性

忠实度的“源信息”通常由输入内容所决定，而事实性的“源信息”则表现为了“证据”，即用于生成过程的常识知识(Commonsense Knowledge). 因此，基于事实性生成的核心问题在于如何确定“证据”的具体形式，并形成事实性的质量控制机制. 根据“证据”信息是否需要进一步的改造和推理，特别是开放式文本生成和反事实文本生成这两种方式已成为目前事实性生成的热点，相关模型如表 4 所示.

表3　忠实度生成模型

对比文献	主要结构	特征	数据集	评估结果			
				输入覆盖比率	幻觉信息比率	BLEU	PARENT
Rebuffel等人[39]	编解码模型	增加词级标签	Wikipedia	—	1.43	41.56	79.00
Rashkin等人[40]	编解码模型	增加句子级标签	Wikipedia	72.30	16.90	8.90	—
Liu等人[42]	BART	文本规划	WIKIPERSON	99.73	0.60	17.12	56.75
Zhang等人[43]	Transformer	句子级Mask	XSUM	43.60	0.26	—	—
Aralikatte等人[44]	Transformer	词级焦点采样	XSUM	44.80	0.26	—	—

3.2.1　开放式文本生成

开放式文本生成不仅需要考虑到上下文之间的因果逻辑,而且需要在生成内容中融合常识知识来构成生成结果的事实性“证据”. 由于现实世界中的常识知识往往具有多类别、多样式和多模态的特征,因此,如何针对开放环境下,结合特定领域知识图谱的富知识语义来进行常识表示,并引导内容的生成则成为该领域的一个关键方向. 本文在Zhu等人[47]工作的基础上将基于知识图谱的文本生成归为以下三种:一是采集知识图谱信息并注入生成模型;二是基于序列化三元组的内容生成;三是带有逻辑关系的事实生成. 具体方法如下.

第一种方法的核心是如何设计一个实现相对独立的知识图谱和生成模型之间的注入策略. Zhou等人[48]以开放式对话任务为背景提出了一个基于静态和动态图注意力机制的CCM模型,该模型从知识库中检索出大量基于常识的图谱信息,通过静态图注意力机制将知识图谱和输入内容融合,再通过动态图注意力机制提升生成内容的准确性. 在基于Reddit改进数据集的测试结果表明,相比仅采用知识增强的MemNet模型[49],CCM模型[48]的困惑度(Perplexity, PPL)值从40.27降低至了39.18. 但由于很难通过知识图谱中的常识知识来直接训练生成模型,因此,如何利用知识图谱中的知识三元组来表示促进生成结果的事实性则是一个重要的研究方向.

第二种方法将知识图谱中富语义的大量三元组直接用于训练生成模型. 例如,Guan等人[50]利用ConceptNet和ATOMIC知识图谱将大量的常识三元组转化为序列化的文本内容,并作为生成模型的重要训练数据以提升生成文本的事实性质量,同时,通过增加事实性分类来提升生成文本的内在因果关系. 在ROCStories数据集的实验验证结果表明,Guan等人[50]的模型在GPT-2模型[12]的基础上,BLEU值从25.7%又进一步提升到了32.6%. 尽管这种模型利用三元组之间存在的因果关系能够提升生成文本的事实性,但是三元组逻辑关系的复杂性与时序动态性,特别是随着知识图谱的扩展以及逻辑词数量的增多,会对事实性文本生成的质量控制产生较大的影响.

第三种方法采用事件序列与关系来表示知识图谱中的复杂逻辑关系. Yu等人[51]提出的CoCoLM模型利用大规模事件知识图谱ASER构建复杂逻辑关系中的事件序列和关系,并以双Transformer编码器为基础分

表4　事实性生成模型

对比文献	主要结构	特征	数据集	评估结果			
				F1 Score	BLEU	Accuracy	PPL
Zhou等人[48]	编解码模型	静态图注意力和动态图注意力	Reddit(modified)	—	—	—	39.18
Guan等人[50]	GPT-2	来自ConceptNet和ATOMIC的三元组, 事实分类器	ROCStories	86.70	32.60	67.07	7.85
Yu等人[51]	Transformer	事件掩蔽机制,事件关系掩蔽机制	ROCStories	—	—	89.40	—
			MATRES	—	—	75.50	—
			COPA	—	—	91.30	—
Madaan等人[52]	Transformer	以条件为导向的文本扰动,四种事实性损失函数	Dbpedia	—	—	72.09	—
			Agnews	—	—	42.76	—
			Yelp	—	—	70.29	—
Calderon等人[55]	T5	针对非枢纽词的破坏和重建	Multi-domain dataset	—	—	81.80	—
			MANtIS	75.40	—	—	—
Paranjape等人[53]	T5	基于知识检索的生成	HotpotQA	53.36	—	—	—
			BioASQ	42.89	—	—	—
			AQA	28.90	—	—	—

别构建了事件预测和逻辑关系预测的掩码语言模型，实现了由逻辑关系控制的内容生成．在基于ROCStories数据集的生成评价中，相比Guan等人[50]的事实性生成模型，CoCoLM模型的Accuracy值再次从67.07%大幅提升到了89.4%．

3.2.2 反事实文本生成

反事实生成是将已有事实视为初始“证据”，并通过检索或推理的方式获取任务领域中的新事实信息来作为“新证据”，使生成的文本内容保持可信度．其本质是基于“证据”进行文本内容的构建，特别是当生成任务缺乏足够的训练数据时，利用反事实生成技术可以促进生成内容的领域适应性．

根据“证据”的具体形式和来源可以将反事实生成技术分为关键词扰动和外部知识引入两类．在基于关键词扰动的反事实生成领域，Madaan等人[52]提出的GYC模型采用关键词和外部数据集实现事实性控制，对情感、主题和时态等关键词构造词级扰动，并通过分类器和奖励模型控制生成文本的事实性质量，实现文本内容的反事实改造．基于Yelp数据集的实验说明，相比采用随机扰动的BERT模型[10]，GYC模型的Accuracy值从10.87%大幅提升到了70.29%．另外，在基于外部数据集的反事实生成方面，Paranjape等人[53]提出的RGF模型采用了基于Wikipedia的内容检索技术，将QA任务中的“问题-答案”对扩展为包含不同领域知识的多个“问题-知识-答案”三元组，再基于匹配程度最佳的三元组生成带有新领域知识的文本内容．基于BioASQ数据集的测试结果表明，相比随机知识增强的Agen-Qgen模型[54]，RGF模型的F1值从32.58%提升到了42.89%．考虑到反事实生成前后的多个“证据”词之间可能存在的矛盾，Calderon等人[55]提出的域迁移模型DoCoGen通过破坏和再重建的方式先剔除了初始域中的“证据”，再结合“被破坏的句子”和新领域的“证据”内容，重建了适应于目标域的反事实文本．在基于多领域迁移数据集的情感分类实验中，相比仅采用目标域数据训练的Transformer模型，DoCoGen模型的Accuracy值从78.8%提升到了81.8%．

3.2.3 小结

开放式文本生成和反事实文本生成的本质都是基于结构化或非结构化“证据”的生成控制和迁移．从利用知识进行开放式生成的视角看，知识图谱的采集和注入方法[48,49]具有两项优势：一方面可以实现领域知识图谱构建和生成模型训练的并行工作，另一方面也可以降低训练难度以提升系统开发效率．但这类模型的生成文本也可能存在预训练数据和知识图谱之间的语义冲突．所以，采用基于序列化三元组的文本生成[50]能够充分优化生成模型对知识的理解和运用，但是三元组规模的扩大和幻觉信息的掺杂都可能影响生成文本的事实性质量．基于知识图谱复杂逻辑关系的文本生成[51]缓解了三元组信息与生成模型中预训练数据之间的潜在语义冲突，也有助于生成模型对知识信息的编码．但现实中基于复杂逻辑关系的知识图谱的生成模型训练是困难的，而且生成模型也会面对难以任务多样化或领域迁移的问题．从反事实文本生成[52,55]的角度出发，这类生成任务解决了目标领域“证据”充足但训练样本不足的困难，通过借助生成质量良好的基座生成模型，可以达到生成模型反事实改造，同时有助于扩增训练数据或无偏数据集的构建[53]，因此可以广泛地应用于特定领域中分类模型的测试[52]．但是反事实生成的内容中仍然可能存在“证据”冲突．如何协调不同“证据”之间的关系，保证整体事实性的质量要求仍然值得进一步研究．

此外，从以ChatGPT为代表的大语言模型的实际表现来看，事实性错误依然没能完全避免，尤其是当用户恶意反馈的错误信息被模型所理解和运用时，大预言模型为了保证与用户的意图对齐可能会生成进一步错误的内容，甚至可能会导致虚假信息的扩散．因此，如何利用客观常识知识来控制或纠正生成文本的事实性质量是未来的一个重要的挑战．

3.3 连贯性

本文基于Halliday等人[38]对连贯性的理解并映射到文本生成中，将连贯性质量特征定义为局部文本或者独立的段落应当服从的某个特定“范畴”(Scope)，如关键词、Memory模块、槽值、计划(Plan)或草稿(Script)等，满足范畴的多个句子的语义之间高度内聚、动态变化且连续．图4为基于连贯性的典型文本生成模型，可以看出，文本连贯性的核心问题就是如何对“范畴”进行编解码并控制连贯性内容，尤其是端到端和多阶段生成模型中的连贯性控制方法，相关模型如表5所示．

3.3.1 端到端生成模型

基于传统端到端模型的生成文本往往存在着语义漂移或语义脱离的现象，这导致了生成文本的连贯性不足．根据“范畴”在端到端模型中的形式，目前解决语义漂移或语义脱离的方法主要归纳为基于关键词的连贯性控制、基于Memory模块的连贯性控制、基于槽值知识的连贯性控制这三种类型．

基于关键词的连贯性控制的基本思想是，通过提取显式或隐式关键词来控制生成文本的连贯性．在提取显式关键词方面，Zhang等人[56]提出的STAR-BTM模型采用相互独立的编码器分别提取历史话题关键词和当前对话的话题关键词(图4(*b*))，并融合为解码器的输入以保证生成内容语义在不同话题关键词之间平滑且连贯．基于JDC数据集上的测试结果说明，相比传统

的编解码模型HRED[57],STAR-BTM模型的BLEU值从12.0%提升到了13.38%. 而在隐式关键词提取方面,He等人[58]提出了SIRM模型,该模型从全文中分别获取粗略和精细的隐式关键词信息(图4(*c*)),该机制有助于识别细粒度的隐藏语义并提高相关生成文本的连贯性. 基于Spam数据集的测试结果表明,相比朴素Transformer模型,SIRM模型的F1值从86%提升到了88.18%.

基于Memory模块的连贯性控制的基本思想是,提取出内容语义、用户行为等不同维度的关键特征来构造Memory模块. 例如,Qin等人[59]提出的CMR模型从对话历史和文档语义知识来进行编码和融合(图4(*d*)),通过BiLSTM获取融合的上下文语义以形成Memory模块[59,60]并用于控制解码器的连贯性输出. Tian等人[61]在CMR模型的基础上进一步提出了RAM模型,该模型采用知识蒸馏的方式构建两种Teacher模型(图4(*g*)),分别用于捕获来自参考文本和输入文本之间的关键词(RTI)以及关键词之间的关系(RPI),而Student模型则受到两种Teacher模型的指导并形成最终的Memory权重矩阵,进一步对Memory权重矩阵进行解码,输出更加准确且连贯的生成文本. 在Reddit数据集上的实验结果表明,RAM模型采用RTI和RPI的F1值在CMR模型的基础上,分别提升到了11%(RTI)和3%(RPI). 此外,Xu等人[62]认为采用GNN网络也可以将用户输入和对话历史进行表示融合,作为一个独立的Memory模块(图4(*e*)),并将对话历史中的每一句话构建为网络中的一个节点,该模型采用GNN网络计算节点的相关性以获得当前对话语义中最相关的节点,用于指导解码器生成与上下文连贯的回复内容.

基于槽值知识的连贯性控制的基本思想是,利用用户意图和槽值在任务域对话中的核心作用,来控制生成文本的动态语义变化并保持连贯性. Peng等人[63]提出的SOLOIST模型集成了四个主要模块:会话理解模块、会话状态跟踪模块、会话策略模块和回复生成模块. 该模型使用了意图词和槽值作为贯穿四个模块的主要线索,并基于自回归模型控制会话上下文中的连贯性(图4(*f*)). 在CamRest676数据集的实验结果表明,SOLOIST模型输出文本的BLEU值在GPT-2模型19%的基础上,进一步提升到了25%.

3.3.2　多步骤生成模型

在长文本连贯性生成任务中,文本计划(Plan)或草稿(Script)取代了关键词成为连贯性“范畴”的具体形式. 为了更加精准地控制生成内容的细节,多步骤的生成模型成为长文本连贯性控制的重要方法. 根据不同步骤所采用功能模块差异,整个模型又分为不同步骤采用不同功能模块、不同步骤使用相同功能模块两种类型.

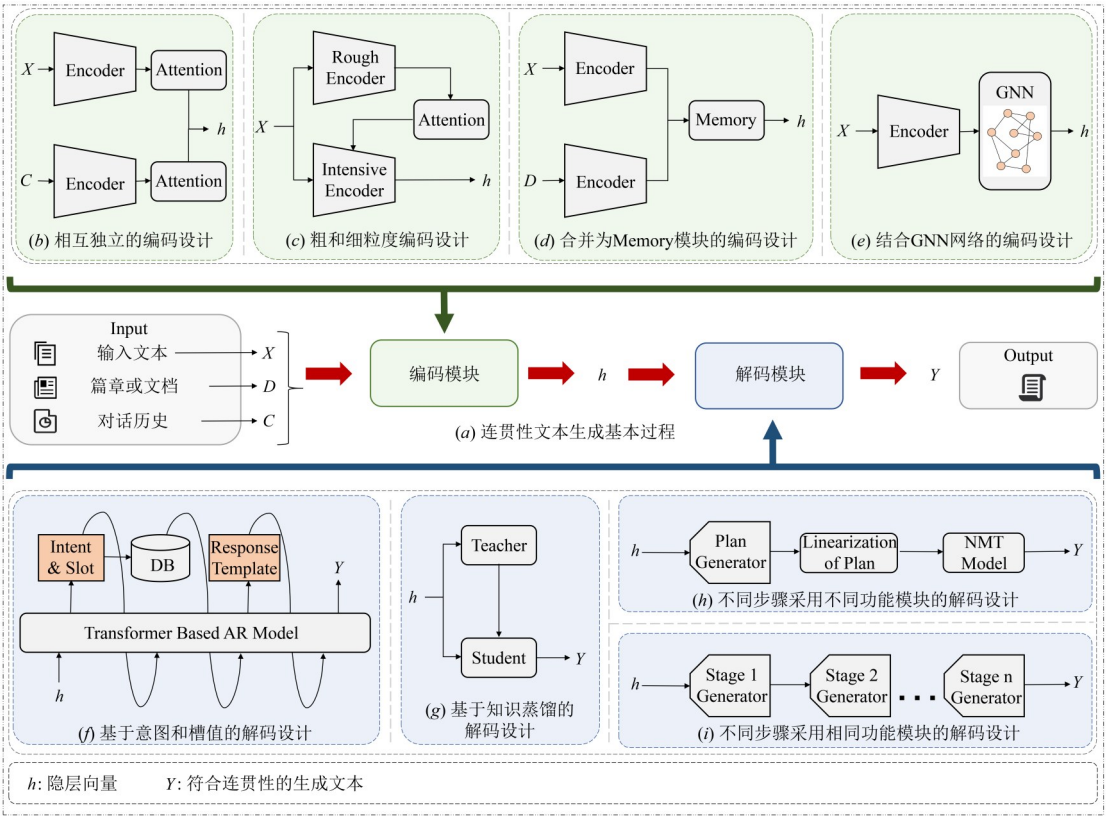
在不同步骤采用不同功能的模型设计中,一般包含两个步骤:第一步是根据多个输入信息生成文本计划或草稿;第二步则通过自回归网络生成包含该计划或草稿的长文本. 例如,Moryossef等人[64]提出的Best-Plan文本生成模型将实体信息构建为有序树(Ordered Tree),然后再将该有序树线性化(Linearization)生成文本计划(Plan),通过机器翻译模块输出为满足连贯性质量要求的长文本(图4(*h*)),实验结果表明BestPlan比基于规则的生成模型UPF-FORGe的BLEU值38.5%高出了8.9%. 由于BestPlan的工作方式类似于选择实体词并构建句子的过程,而Hua等人[19]提出的PAIR模型更像是对生成的计划进行完形填空[65]. PAIR采用基于BERT[10]和基于BART[4]的2种文本计划生成器来实现了文本草稿(Script)的生成与构建,从而规避了构建实体词与实体词之间复杂关系的困难. 在NYT(NEWS)数据集的测试中,相比标准的BART模型,PAIR模型的BLEU值从11.6%提升至34.3%. 从上面两种方法的讨论中可以发现,在不同阶段采用不同模块的机制有两个优点:一是每个步骤都可以单独训练以满足预期结果的要求,从而保证了模块化构建和持续优化;二是由于具有独立的计划生成器,从而可以结合特定目的和策略,直接影响到了后续文本的生成效果.

在不同步骤采用相同功能模块的模型设计中,由于每个步骤使用的模块不需要单独设计,因此仅存在训练数据的差异. 例如,Tan等人[18]提出的ProGen模型在每个步骤中使用相似的BART模型[4]作为子模块并提供不同的触发词,这些触发词构成了文本计划(Plan)(图4(*i*)). 由于每个子模块可以独立地通过触发词进行训练,所以可以得到基于触发词精确控制的连贯长文本. 实验结果显示,ProGen比标准的BART模型的BLEU值30.1%提高了1.1%. 由于ProGen模型每个步骤的目标不同,每个子模块获取的训练数据集也不同,虽然不同的训练数据集保证了每个子模块稳定有效的训练,但在实际中构建这些数据集也存在一定的困难.

3.3.3　小结

连贯性文本生成的关键思想就是将核心语义贯穿于文本生成的控制过程中,使其能够符合统一的“范畴”. 基于关键词[56,58]、Memory模块[61,62]和槽值知识[63]的连贯性文本生成具有功能模块相对独立和易控制的优势,生成文本的前后文过度平滑且关联性较好. 但也需要针对性地实现不同类型文本或不同粒度语义下的融合设计,以及为了防止语义漂移现象出现的解码器注意力设计. 然而,Marchenko等人[66]仍然认为大部分

采用端到端结构的自回归模型并没有真正实现连贯性的质量要求，因为关键语义信息可能在生成文本逐渐增长的过程中出现损失，并导致连贯性缺失，所以采用多步骤控制的连贯性文本生成被用于长文本生成和故事续写任务. 其中，不同步骤采用不同功能的模型[19,64]保持了模块化设计和独立训练的特质，使其能够提取出文本控制的“计划”或“草稿”，以增强生成过程的可解释性. 而在不同步骤采用相同功能模块的模型中[18]，通过设定不同触发词以控制生成过程，与采用CoT[67]构造提示信息的大语言模型生成具有相似的思想. 此外，值得注意的是，大语言模型的出现意味着生成文本需要符合更高层级的连贯性需求，尤其是以思维链和思维树[68]为代表的逻辑连贯性显著地提升了推理能力，并直接影响了生成结果，所以如何通过构造更加精细的提示信息以实现更加细致的文本连贯性控制值得长期研究.



注：(*a*)表示此类模型的基本生成过程，(*b*)~(*d*)均采用多编码器设计实现不同类型信息的独立编码、粒度控制、信息融合，(*e*)通过GNN网络构建连贯性文本之间的关系，(*f*)和(*g*)通过槽值和教师模型约束连贯性内容生成，(*h*)和(*i*)通过Plan和Script控制多步骤文本生成中的连贯性.

图4 基于连贯性的典型生成模型

3.4 一致性

文本一致性关注于生成文本与上下文之间的静态“属性”是否一致. 其中，角色属性的一致性有助于减轻生成内容中角色差异对用户造成的困扰，而情感属性一致性有助于提升生成文本中的感性表达. 图5表示基于角色一致性的典型生成模型，研究文本一致性的核心问题在于：如何确保静态“属性”在生成文本中保持不变. 本节对基于角色和基于情感两种属性的一致性模型设计方法分别进行讨论，相关模型如表6所示.

3.4.1 角色属性的一致性

一般地，角色属性的显式表达以角色的属性模型为主要形式，而隐式表达则由该角色说过的话、相关行为记录所构成. 当前研究存在着两方面问题的影响：一是现有的训练数据集中往往存在着大量与角色无关的噪声信息，导致生成内容的角色属性表达不够鲜明或存在着偏差；二是生成文本难以同时兼顾语义内容、角色属性以及与角色相匹配的知识. 因此，基于角色属性一致性的研究思路主要包括以下三种：针对角色属性的数据增强；针对语义、角色和知识的编码控制；针对角色强化的解码机制.

针对角色属性的数据增强方法的本质思想是通过构建无偏数据集来提升模型对角色属性的编码质量，并促进生成文本角色的属性一致性. 目前主流的方法包括两种. 一种是直接在原始数据集上进行编辑. 如Cao等人[69]提出的D3模型通过删减训练数据中的冗余

表5　连贯性生成模型

对比文献	主要结构	特征	数据集	评估结果			
				F1 Score	BLEU	ROUGE	METEOR
Zhang等人[56]	编码器-解码器,注意力机制	主题注意力模型,联合解码器	JD contest	—	13.38	—	—
			Ubuntu conversation dataset	—	13.30	—	—
He等人[58]	编码器,GAN	粗略和精细理解模型	Reddit/movies	70.01	—	—	—
			Tweets/ghosh	82.54	—	—	—
			IAC/v1	63.01	—	—	—
			Industry/spam	88.18	—	—	—
Qin等人[59]	LSTM,注意力机制	Memory矩阵	Reddit	38.00	1.38	—	7.46
Tian等人[61]	编码器-解码器,注意力机制	知识蒸馏		49(RTI)	1.43	—	7.74
				41(RPI)	1.40	—	7.59
Peng等人[61]	Transformer	意图,槽值	CamRest676	—	25.50	—	—
			MultiWOZ	—	16.54	—	—
Xu等人[62]	编码器-解码器	GNN	Weibo	—	58.80	—	—
			Douban	—	52.50	—	—
Moryossef等人[64]	多步骤生成	RDF三元组,序列树	WebNLG	—	47.40	63.10	39.10
Tan等人[18]	多步骤生成	多BART模型	CNN News	—	31.10	—	—
Hua等人[19]	多步骤生成	计划(Plan)和迭代微调	Reddit(CMV)	—	36.09	56.86	33.30
			NYT(opinion)	—	23.12	40.53	24.73
			NYT(NEWS)	—	34.37	51.10	29.50

句子,再利用实体替换、数据匹配和角色编辑扩增样本数量(图5(*b*)). 基于PersonaChat数据集的测试表明,相比朴素的Transformer模型[9],采用数据增强后的D3模型将BLEU值从3.14%提升到了3.35%. 另一种是采用预训练语料来增强原始数据集. 如Kim等人[70]利用蕴含先验知识的GPT-2模型构建对偶数据集,并训练教师模型为增强后的角色匹配分值(图5(*f*)). 基于对偶数据集的测试结果表明,相比朴素的Transformer模型,GPT-2模型的METEOR值从10%提升到了20%. 上述两种角色的数据增强方法也都存在一定的缺陷,特别是直接在原始数据集上进行编辑的方法有可能会引入不合理的实体或不匹配的数据,而采用预训练语料去丰富数据集的方法也需要重新评估新构建的角色属性,否则也可能会引入噪声.

针对语义、角色和知识的控制问题,一般可以通过构建相互独立的编解码模块或依存结构模型来解决. Jang等人[20]分别对角色属性和相关知识进行独立编码以创建上下文表示(图5(*e*)),并指导生成内容中的人格和知识的选择. 基于FoCus数据集的实验结果表明,相比朴素的BART模型[4],Jang等人的模型[20]的BLEU值从23.73%提升到了28.68%. 此外,构建依存结构的编码模型能够进一步控制文本语义、角色和知识之间的关联程度. Fu等人[71]采用5个基于BERT模型[10]的编码模块,分别获取角色和知识的先验和后验分布,确保了角色和知识之间的匹配. 而Yang等人[72]基于强化学习设计了一个价格谈判模型(图5(*g*)),该模型融合谈判状态和用户行为构建了用户角色编码,并通过用户角色、谈判状态和行为之间的依存关系实现了基于心智理论的回复生成.

针对角色强化的解码机制,可以通过构造多解码器组合或多解码子任务来实现. Song等人[73]提出了BOB模型(图5(*d*)),它使用两个基于BERT的独立解码器分别控制语义和角色属性,即一个解码器仅用于生成带有基本语义和少量角色属性的文本草稿,而另一个解码器在文本草稿的基础上进一步强化角色属性以确保角色属性的一致性. 值得注意的是:在多人对话任务(Multi-Party Conversation,MPC)中对角色识别有也明确的要求,即需要解决“谁说给谁”的问题. Gu等人[74]提出的MPC-BERT模型通过设计特定的解码器结构来构建“说话人搜索”子任务,有效增强了生成内容角色的一致性.

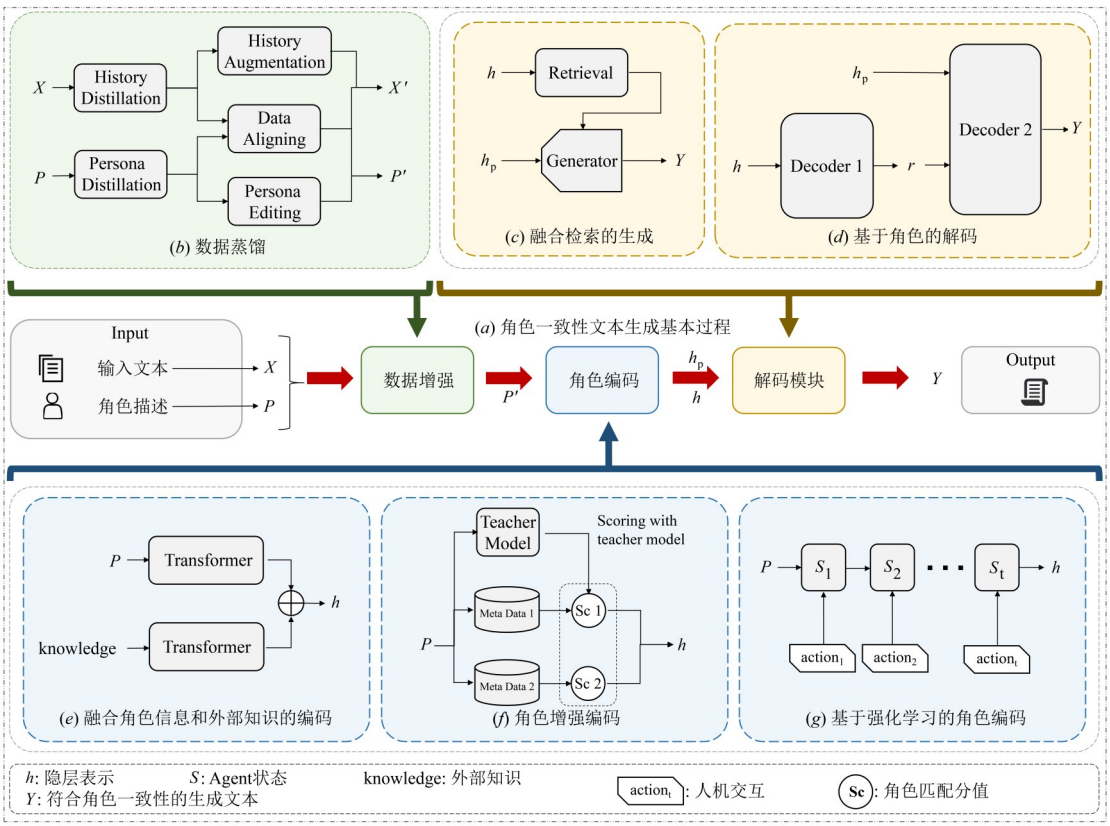
3.4.2　情感属性的一致性

情感属性的一致性常被用于阻止生成文本中出现的情感错误或者情感突变,其核心任务是通过构造情感向量来控制上下文中的情感属性. 一般常采用两种方法. 一种是基于角色特征来预测说话人的情感向量. Wen等人[26]提出了一种融合对话历史和角色特征编码的PET-CLS模型用来捕获情感变化,然后组合历史情感和当前情感的变化来预测未来情感. 在PELD数据集的实验结果表明,相比基于RoBERTa[13]编码的情感预测模型,PET-CLS模型[26]获得了更高的性能,其情感

一致性F1值从RoBERTa的38.9%提升到了43.1%. 另一种方法是依赖历史情感的连续性来控制当前的情感向量. Mo等人[25]提出的GMTF模型计算对话历史中每句话的情感向量并组合为情感矩阵,矩阵中记录的情感变化促进了生成文本的一致性情感预测. 基于ROC-Stories数据集的测试结果表明,相比朴素的Transformer模型[9],GMTF模型[25]的BLEU值从23.77%大幅提高到了27.03%. 因此,采用情感的历史特征来约束和控制情感的变化行为则是保持情感属性一致性的重要策略.

3.4.3 小结

从生成文本中的静态"属性"前后一致的角度出发,本文讨论了角色一致性和情感一致性. 角色一致性有助于生成模型捕获说话者和接收者的内核以提升生成内容的合理性,所以采用角色数据增强的方法的本质就是为了强化角色表达使其内核更加鲜明[69,70]. 但无论是通过直接编辑数据样本还是利用预训练模型的先验知识来构造无偏数据集,都无法彻底避免潜在噪声. 如何构造可靠的角色评估模型也是一个关键的难点. 角色属性还可能包含角色所适配的知识和相关的语义信息,所以采用属性间的独立或相互依存编码机制是必要的[20,72]. 但是如何确保这些相关属性间的相容性以克服潜在的语义冲突则是另一个难点. 值得注意的是,大预言模型中的角色属性控制将会更加细致且复杂,包括角色的性格偏好、行为方式、价值观念、角色和事物之间的关系、角色和角色之间的关系等. 这些不同属性和关系的建立以及相容性都将影响不同智能体角色[75,76]的一致性质量. 如何组织和编码上述属性以形成综合性的角色内核成为当前研究的一个热点. 此外,在多个角色之间的协作和交流中,通过特定解码器设计和多解码器组合机制[73,74],能够提升角色表达的准确度或针对性. 但是如何基于心智理论(ToM)以及马斯洛层次需求(Maslow's Hierarchy of Needs)等理论体系,增强角色对其他角色的认知层面的理解[77]应当成为未来的角色一致性研究的目标之一. 从情感一致性的角度看,现有方法在预测生成文本的情感属性时,侧重于依赖与角色表达和历史情感变迁[25,26],这有助于保持情感的连续性并提供一定的可解释性. 但是这种方法忽视了现实中情感变化所依赖的多种因素,包括生理、环境、事件和思维方式等. 所以如何利用大语言模型将多种因素纳入情感一致性的控制过程值得进一步研究.



注:(*a*)表示此类模型的基本生成过程,(*b*)基于数据蒸馏实现角色信息的数据增强,(*c*)将检索信息融入内容生成,(*d*)通过不同解码器控制语义和角色,(*e*)表示角色和知识的编码融合,(*f*)利用教师模型控制角色增强,(*g*)基于Agent状态进行角色编码.

图5 基于角色属性一致性的典型生成模型

表6　角色一致性生成模型

对比文献	主要结构	特征	数据集	评价结果				
				F1 score	PPL	BLEU	ROUGE	METEOR
Cao等人[69]	Transformr或GPT-2	针对角色信息,回复信息和对话历史的数据蒸馏和编辑	PersonaChat	—	37.30	3.35 (Transformer)	—	—
				—	15.69	4.18 (GPT-2)	—	—
Kim等人[70]	双编码器,GPT-2	对偶任务数据集构建,角色表达排序	PersonaLink	—	—	—	—	20.00
Jang等人[20]	GPT-2或BART	独立编码器组合	FoCus	—	11.45	15.21 (GPT-2)	27.36 (GPT-2)	—
				—	23.25	16.18 (BART)	28.68 (BART)	—
Wen等人[26]	RoBERTa	VAD向量,角色	PELD	43.10	—	—	—	—
Mo等人[25]	Transformr	情感向量矩阵	ROCStories	—	—	27.03	—	—

3.5　多样性

文本多样性生成可以理解为通过不同的风格或样式来表达相同的文本语义. 如图6所示的多样性典型生成模型,多样性质量特征在文本风格迁移任务中主要表现为风格编码,而在其他任务中也可能被表达为领域编码、焦点或内容选择等. 本文将从基于风格的多样性模型和其他多样性模型两个方面进行讨论,相关模型如表7所示.

3.5.1　风格多样性

原始文本中语义和风格相互交错,改变原始风格的过程中可能存在核心语义的损失,所以风格多样性的主要研究问题就是如何精准控制生成文本的风格属性. 现有的研究主要分为两种思路:一是在分别提取原始文本中的语义和风格,即实现风格与内容特征解缠的基础上,来生成满足特定风格要求的文本;二是直接在原始文本的基础上增强目标风格的表达.

基于第一种思路,根据风格提取方法又可以分为两种. 第一种是直接对原始文本中的词或词组进行编辑,构建无风格的不完整文本,并以此为基础再生成具有目标风格的新文本. Lee等人[78]基于该方法设计了一种两阶段的SST模型,其中,阶段一通过风格分类器剔除原始文本中的风格词,阶段二通过文本重建损失和风格迁移损失进行质量控制来生成目标风格的新文本. Zhou等人[24]在上述两种损失函数的基础上进一步使用了层级关联传播策略来控制词级风格,并增加了风格关联一致性损失和流畅性建模损失. 在基于Yelp数据集的测试中,该模型在SST模型的BLEU值24.93%的基础上大幅度提升到了60.4%. 第二种方法是通过多个编码器分别提取原始文本中的语义和风格,再通过解码器和风格分类器控制文本质量. 例如,Shu等人[79]设计了FVN模型,该模型使用多个编码器和码本(codebook)模块实现内容和细粒度风格的解缠,并通过复用编码器来控制生成的内容(图7(d)). 在PersonageNLG数据集上的实验表明,相比CVAE模型[80]的BLEU值90%,FVN模型的BLEU值提升了6.5%.

基于第二种思路,增强生成文本中的目标风格,即采用两次风格迁移去控制风格属性的转换:第一次将原始文本转化为带有目标风格的生成文本,第二次是将生成文本通过原始风格再次转化为原始文本. Dai等人[81]采用上述思想,基于Transformer[9]的生成器模型,采用条件风格鉴别器和多风格分类鉴别器去控制转换后的文本风格属性(图7(h)). 而两次风格转换中的模块复用也是重要的改进方向. Cheng等人[82]提出的CAST模型在两次风格转换中选择使用相同的编码器和解码器(图7(f)),并通过增加平行数据训练模块实现了平行数据和非平行数据的同步训练以达到风格控制的目的. 虽然模块复用的设计思想的确可以增强模型的训练,但也会影响实际的风格控制效果,因为两次风格迁移的训练可能对相同模块产生了相互矛盾的影响. 基于GYAFC数据集的测试表明,相比朴素Transformer模型的BLEU值24.16%,CAST模型的BLEU值提升到了26.38%.

3.5.2　其他多样性

与风格多样性不同,其他文本多样性生成的目标是在保证原文本语义不变的情况下,扩增内容和表现形式来促进生成文本的多种表达,而无须满足某种特定的样式或属性. 在对话任务中,增强多样性的基本方法包括了数据增强和离散采样两种. 其中,数据增强方法的本质就是增加训练数据中的回复文本的种类. Khayrallah等人[83]提出了一种通过释义生成的方式将训练样本的参考句子改造成了多种不同表达形式的SMRT模型,其中释义生成器构成了教师模型,并通过

知识蒸馏的方式训练学生模型以生成更加多样性的文本. 在DailyDialog数据集上的实验结果表明:相较于朴素的Transformer模型,经过释义增强训练数据后的SMRT模型的BLEU值从10%提升到了12.4%. 另外,回译技术也可被用于训练集的数据增强. Su等人[84]提出的模型采用回译技术将非对话文本转化为训练数据,由于非对话文本来源更加广泛且蕴含更多信息,所以增强后的训练数据可以有效提升生成文本的多样性. 基于Weibo数据集的实验表明,相比CVAE[80]的BLEU值17.1%,采用回译技术增强后模型[84]的BLEU值提升了0.9%. 而离散采样方法的核心思想是在满足回复内容的邻近语义空间中,随机采样隐向量并解码为具有多样性的生成文本. Hu等人[85]的工作利用VAE模型[86]从隐向量高斯分布中采样并解码实现多样化的表达能力,从而提取出输入内容中的非结构化信息,并结合结构化编码共同实现多样化的质量目标(图7(e)). 在基于IMDB数据集的验证结果表明,相较于VAE模型的Accuracy值62.5%,Hu等人[85]的模型考虑到了非结构与结构信息的共同编码能力,使模型的性能提升到了64%.

此外,多编码器设计和带有焦点引导的解码器改进也有助于增强生成文本的多样化. 针对多编码器设计,Lachaux等人[87]将编码器设计为两个平行结构(图7(b)),一个用于编码原始文本的语义,另一个用于编码特定领域信息. 通过内容和领域信息的分开编码,可以控制领域编码以实现领域多样性. 而针对带有焦点引导的解码器改进策略上,Cui等人[88]提出的解码器使用聚焦约束注意力机制(图7(c)),使编码器网络能够捕获当前上下文主题与响应之间的焦点信号并引导生成. Wang等人[89]提出的模型主要用于问题生成任务,该模型基于上下文和响应信息进行文本生成(图7(g)),并设计了一个内容选择器来提取目标所关注事项特征,从而促进模型生成的文本具有更好的多样性.

3.5.3 小结

在风格多样性中,第一种方法的优势在于能够从原始文本中分离语义和风格,以实现细致的风格控制和细粒度的语义处理,但这种方法也可能会导致特征提取不够彻底或语义损失. 第二种方法的优势在于能够尽量保持原始文本中的完整语义并通过两次风格转换提升模型对风格的理解能力,但这种方法也会存在着风格转换不完全和风格混乱[82]的问题. 此外,如何实现多种风格组合的多样性精准控制也成为未来的重要挑战. 在其他多样性中,数据增强利用释义生成和回译方法降低了文本多样性的训练成本,而离散采样能够在隐向量空间中捕获语义信息的同时增加更多样式的文本表达,但这两种方法也可能引入噪声并导致语义偏离. 多编码器和带有焦点的解码器能够提取原始语义,同时控制目标领域或焦点引导的内容生成. 但也需要面对语义损失和领域适应性的问题.

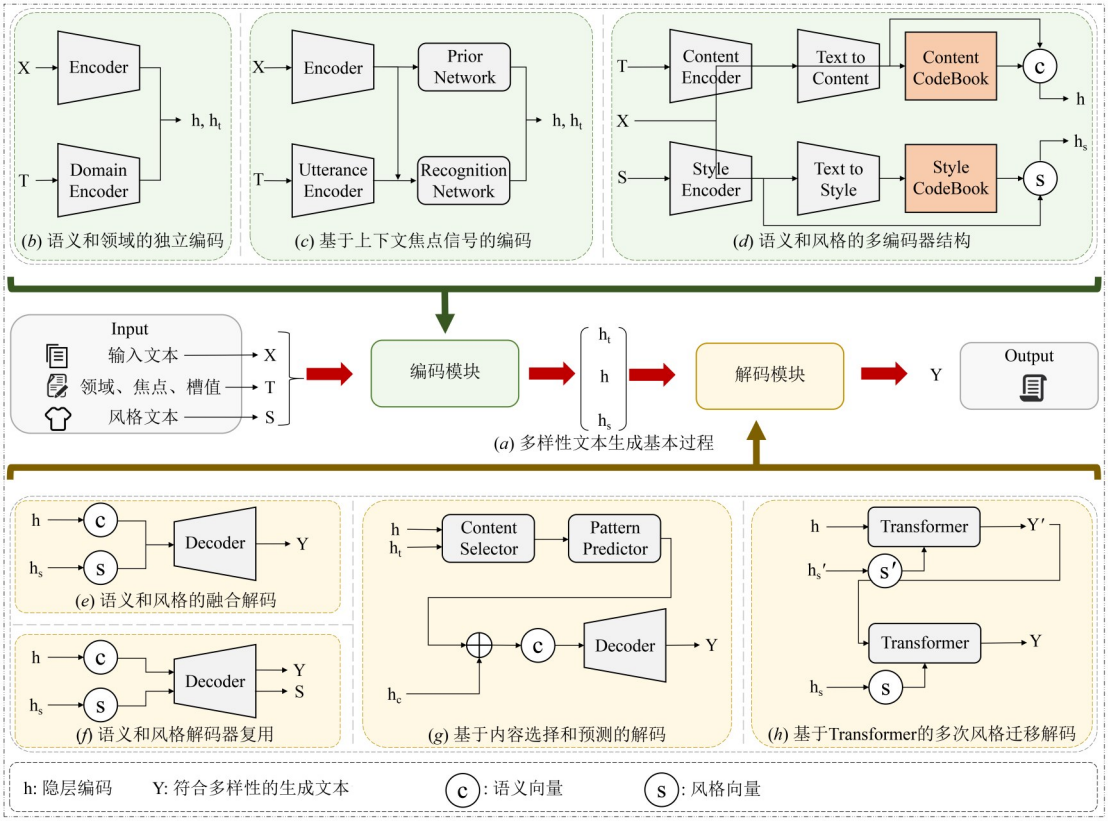
通过ChatGPT等模型[1,2]的测试可以发现,其多样化表达能力相较于其他语言模型[4,12]已经大幅提升,而如何将这种多样化能力赋予中小规模生成模型,来实现特定领域的精准、轻量化以及自适应能力将成为未来研究的工作之一.

3.6 流畅度

当前的流畅度生成模型依然主要存在两个方面的缺陷:一方面,在长文本生成或者持续性的文本生成任务中,生成的内容过长而可能会导致缺少逻辑而陷入语义的混乱,使可读性变差. 另一方面,训练较好的生成模型虽然保持了较好的流畅度和可读性,但关键信息的缺失可能会导致内容语义稀疏而失去语言功能价值. 本节将围绕文本生成过程中流畅度模型涉及的这两个问题进行分析.

3.6.1 长文本流畅度

长文本生成作为生成领域中的重要任务具有广泛的应用价值,它要求保证长距离文本持续生成的同时依然具有稳定的可读性. 而以GPT-3模型为代表的LLM模型在生成较长文本后仍然可能出现语义混乱的现象. 本质原因在于LLM所依赖的Transformer解码器在处理长文本时,会将文本内容切分为多个长度固定的“段”(Segment),导致生成文本更加聚焦段内的局部流畅度. 所以,长文本内容流畅性的关键在于如何将已生成内容的核心语义映射到在后续生成文本的“段”中. Dai等人[16]提出了一种采用段级循环机制的TransformerXL改进模型,该模型使每一个段的编码均基于上一个段中每一层的输出向量,同时,结合相对位置编码克服了段之间独立编码存在的上下文语义不足的缺陷,增强了生成文本对长距离语义的注意力,在One Billion Word数据集上实验验证结果表明,该模型的PPL值比标准Transformer下降了2.2. 此外,XLNet[17]也采用了TransformerXL模型的注意力段级循环和相对位置编码技术,并通过构建内容流和查询流相结合的机制实现了和BERT模型中掩蔽语言模型子任务(Mask Language Model,MLM)相同的效果. XLNet融合了自回归和自编码模型的优点,不仅进一步增强了对上下文语义编码的注意力,而且克服了BERT模型中序列长度不能超过512个词的限制. 实验结果表明,XLNet在RACE数据集上的Accuracy值比BERT模型提升了13.4%,且在SQuAD2.0数据集上F1值比TransformerXL提升了1.74%.



注:(*a*)表示此类模型的基本生成过程,(*b*)~(*d*)采用多编码器设计实现了独立编码、焦点控制、风格控制,(*e*)基于语义和风格控制解码内容,(*f*)基于解码器复用去控制语义和样式,(*g*)通过内容选择和预测控制解码结果,(*h*)采用多次解码控制风格.

图6　基于多样性的典型文本生成模型

表7　多样性生成模型

对比文献	主要结构	特征	数据集	评价结果			
				F1	BLEU	ROUGE	METEOR
Shu等人[79]	编码器-解码器	码本	PersonageNLG	99.40	96.50	76.80	48.60
			E2E	93.50	71.40	71.90	45.10
Cheng等人[82]	编码器-解码器,GAN	平行及非平行数据训练	GYAFC	—	26.38	—	—
			Reddit	—	23.92	—	—
Dai等人[81]	Transformer,GAN	三种子任务	IMDB	—	70.50	—	—
			Yelp	—	54.90	—	—
Zhou等人[24]	编码器-解码器	词级风格相关性		94.00	60.40	—	—
			GYAFC	81.40	47.70	—	—
Cui等人[88]	编码器-解码器	焦点约束注意力	Weibo	—	30.32	—	—
Su等人[84]	编码器-解码器	回译		—	18.00	—	—
			Douban	—	27.40	—	—
Khayrallah等人[83]	Transformer	知识蒸馏	Conversational Data	—	13.80	—	—
Wang等人[89]	编码器-解码器	内容选择	SQuAD	—	15.34	—	—
			NewsQA	—	9.90	—	—
Lachaux等人[87]	编码器-解码器	源编码器,目标编码器	WMT17 (En-De)	—	55.40	—	—
			WMT14(En-Fr)	—	65.90	—	—
			WMT17(Zh-En)	—	34.70	—	—

3.6.2 信息量

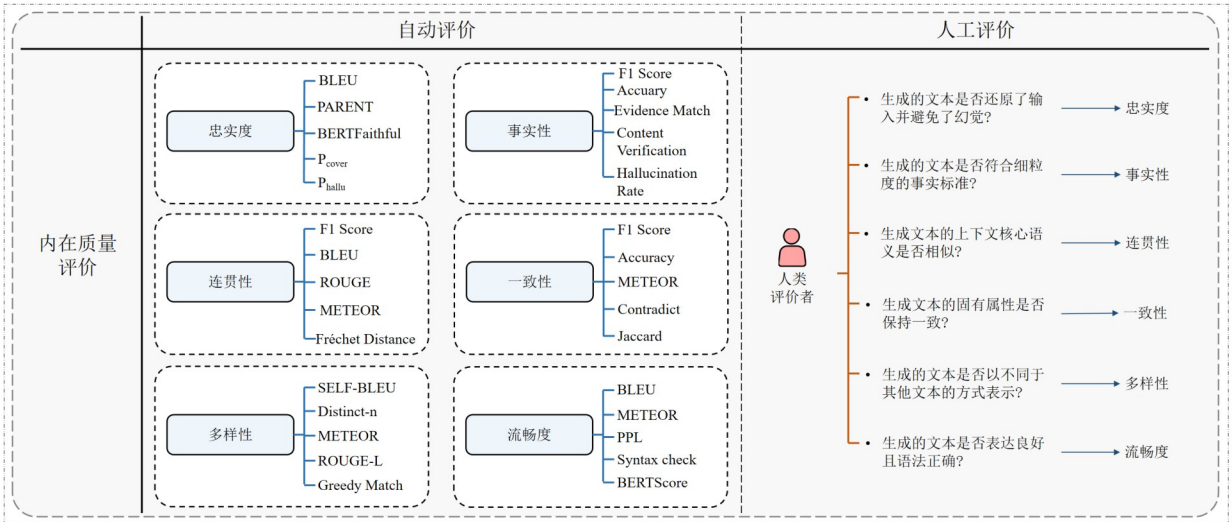
信息量可以理解为生成文本中围绕核心语义的实体词和属性信息的数量. 生成模型为了保持局部流畅度并降低错误率偏向于生成高概率的常用词,导致信息量的下降,所以如何从数据中获取更多用于文本生成的信息十分重要. Sun 等人[90]提出的模型基于 RoBERTa 模型提出一种句子级策略识别模型,并利用心理学领域的特定数据集进行训练,以获取更为精确和细粒度属性的内容信息. 此外,通过摘要生成来捕捉上下文关键信息也是重要的方法. Xu 等人[91]基于 Transformer 模型[9]构建了 SumMem 模型并发布了 Multi-Session Chat 数据集,通过对每一个会话(session)生成摘要的方式记录会话内容的细节信息,再通过检索增强的方式促进生成文本包含更加充分的信息量. Gu 等人[92]提出的 MemSum 模型使用多步马尔可夫决策过程,通过局部句子编码器、全局上下文编码器、历史编码器动态地提取不同视角下的语义信息并形成摘要,同时保证生成文本的流畅度.

3.6.3 小结

相比多样性,流畅度是生成文本中较为基础的质量特征,所以相关研究主要聚焦于生成内容的可读性和信息量. 在长文本流畅度方面,现有方法采用段级循环和相对位置编码等技术,将核心语义映射到后续生成文本的“段”中,这种加强段级注意力的方法促进了生成文本的上下文可读性[16,17]. 在生成文本的信息量方面,要求生成文本保持流畅度的同时尽量回避语义信息弱的常用词,相关模型设计思想类似于忠实度和连贯性对具有信息量的实体和属性词的控制,但流畅度更加侧重于在保证文本可读性的基础上,进一步利用检索、编码和语义增强去促进信息词的表达[91,92]. 事实上,大语言模型的出现已经极大地改善了生成文本的流畅度,甚至在教育、法律和学术研究领域能够表现出比人类写作者更加非凡的语言能力. 但是受大语言模型在训练开销、部署成本和使用价格方面的约束,提升中小语言模型的流畅度尤其是长文本的生成质量对工业界仍然具有重要意义.

4 文本内在质量评价研究现状分析

尽管 Celikyilmaz 等人[32]将文本评价方法分为人工评价、非训练的自动评价算法和评价模型三类,但是考虑到大部分的研究将非训练的评价算法和评价模型联合使用,同时,针对不同的质量特征和需求,生成文本的评价也存在不同的侧重,本文从文本内在和外在质量出发,将评价方法统一到自动评价和人工评价的整体框架中,如图 7 所示.



注:针对每种质量特征的评价机制,由多种自动评价方法和特定人工评价思想组成.

图 7 基于不同质量特征的文本评价框架

4.1 自动评价

忠实度评价的核心思想是计算生成文本对输入内容的覆盖率以及幻觉词[36]在生成文本中的比率,具体评价方法可以分为 5 类,如表 8 所示. 基于传统词重叠的质量评价方法适用于大部分生成任务,如 BLEU 值[93]等,但词重叠评价依赖可靠的数据集而且被认为和人类评价结果之间存在明显差异[94],这是因为参考文本中少量的幻觉信息可能被模型学习并影响基于词重叠的评价结果. 因此,计算生成文本和输入内容之间的差异常被用于改进忠实度的评价方法,如 PARENT[95]利用精确度、召回率等方法来计算数据到文本生成任务中的生成质量. 基于生成词比率的忠实度评价被广泛

使用,即通过计算生成内容中覆盖输入文本的实体词或属性的数量以计算覆盖率[42,43],以及通过计算生成内容中幻觉词与生成文本长度的比值以计算幻觉比率等方法[39]. 此外,Aralikatte等人[44]引入的BERTFaithful指标通过计算测试集中的忠实样本占比去衡量整体忠实度. 基于问答机制的评价思路是从生成文本出发,通过QG模块生成特定问题,再通过QA模块依靠输入文本或文档进行回答,回答内容和生成文本之间的匹配分数被用作忠实度的评估值[96]. QuestEval[97]在上述方法基础上增加了从数据样本中提取问答对,并基于生成文本回答问题以判定忠实度. 基于NLI的评价方法也得到了深入的探索,即通过将输入文本和生成文本拼接,由NLI模型给出评价类别或分值[98],NLI模型[99,100]的训练数据集一般包括MNLI,ANLI,DECODE等. 此外,为了获得更精准的评价,Goyal等人[101]设计了DAE方法,即通过顺序地预测生成文本中每个依赖弧的蕴涵得分以获得最终的忠实得分.

事实性评价的核心思想是如何计算生成文本和"证据"之间的匹配程度,具体分为5种评价类型,如表9所示. 其中,基于比率的统计方法通过计算事实性句子的占比来判定生成模型的性能,包括F1值和Accuracy值等指标已经成为事实性研究工作[50,53]的基本组成. 基于分类的评价将生成文本和证据之间的匹配任务转化为分类任务,一般包括"证实、无关、矛盾"三类标签. Kryscinski等人[102]通过构建"句子-证据"对,利用基于BERT的分类器确定生成内容是否与"证据"语义匹配以及矛盾内容的范围. Dziri等人[103]在此基础上,针对基于知识的对话生成任务评价,通过人工注释的方式建立增强数据集并训练分类器以实现事实性分类. 与忠实度的评价方式相似,事实性同样采用了基于QG和QA模块的问题生成和答案匹配机制[104]. Nan等人[105]从生成文本出发,设计了同时构造问题和答案对的语言模型,简化了基于问答机制的评价方法;在基于NLI的评价时需要面对现有数据集无法训练事实性评价模型的困难,因此需要针对性地人工标注以构造事实性数据样本. Qin等人[106]提出的事实性评价方法CI-ToD使用了基于人工标注的NLI标签,使NLI模型能够判断生成的回复内容是否存在事实性矛盾,而且还提供了更细粒度的分类(对话历史不一致、用户查询不一致或知识库不一致). 而Gupta等人[107]进一步设计了包括待检测内容识别、证据检索和基于NLI模型的事实检测管道DialFact. 最后,考虑到事实性评价的一个关键难点是如何从海量的开放式知识语料中定位相关"证据",Lee等人[108]将Wikipedia设置为"证据"的主要来源并分为了文档级证据和句子级证据,分别对应于幻觉实体度量和内容证实度量两个指标. 幻觉实体度量是指采用文档级的实体匹配方法查找生成文本中的幻觉词,内容证实度量则通过计算生成文本和句子级证据的相关性确定生成内容是否被证实.

连贯性评价的核心思想是如何计算生成文本中多个句子之间的动态语义相似性或关联程度,如表10所示. 由于以BLEU[93],ROUGE[109],METEOR[110]为代表的词重叠评价方法[93,111]计算简单而被大量连贯性论文使用,并且考虑到关键词对连贯性语义控制的重要性,Qin等人[59]通过F1值计算关键词在生成文本和文档中的匹配程度,作为生成文本和文档之间的连贯性分值. 语义相似性对文本连贯性要求更高. Tan等人[18]采用Fréchet Distance指标提取不同文本内容编码后的浅层、中层和深层特征信息,并计算文本编码信息之间的差异作为语义连贯性的分值. Xu等人[112]通过编码两两相邻句子之间的语义向量以及前馈神经网络,获得相邻句子之间的连贯性分值,而所有相邻句子之间连贯性分值的平均值被用于衡量整体生成文本的连贯性. 此外,基于实体的连贯性评价在长文本生成和故事续写任务中十分重要. Elsner等人[113]认为实体词的代词共指是文本连贯性的一个重要方面,并通过计算前后文中不同代词共指的概率来判定文本的连贯性以及范围. Roemmele等人[114]以故事生成任务为背景,同样采用计算生成内容和背景故事之间的实体词的共指概率,计算生成文本的连贯性. 除了实体词,由具有相同词干的重复词构成的词链也可以被用于评价连贯性,Gong等人[115]认为生成文本中通常存在多条词链,通过在机器翻译数据样本中加入词链标签,并计算生成文本词链和参考文本词链之间的匹配值,可以得到整体连贯性分值.

一致性评价的核心思想是如何计算生成文本中静态"属性"的相似性,它包含了特定属性的相似性以及不同属性的匹配度,如表11所示. 基于词重叠的评价方法被广泛用于词级语义一致性的评价,如METEOR通过计算生成文本和参考文本之间的准确率和召回率的调和平均值,实现了关键语义中的属性相似度. 基于一致性内容的比率也被用于衡量生成质量. 在角色属性一致性方面,Kim等人[70]在以角色属性为基础的生成任务中,采用Contradict指标判断前后文本的角色属性是否一致,它的本质是计算回复内容与角色属性的矛盾内容的比例,即角色属性的一致性错误率. Jang等人[20]主要考虑角色属性C_p和知识C_k之间是否匹配,其中C_p由五个给定的角色句子组成,而C_k包含了维基百科的基本事实句子,最终通过Accuracy计算C_p和C_k与生成文本之间的差距. 在情感一致性方面,Wen等人[26]采用宏平均(Macro Avg)和加权平均(Weighted Avg)评价预测的情感和历史情感之间的一致性,较高的宏平均意味着测试集中所有类型情感的一致性较好,而较

高的加权平均意味着测试集中出现类型最多的情感的一致性较好. 基于词级相似性和语义相似性的评价方法同样适用于一致性的评价. 例如Jaccard相似性可以被用于度量生成文本上下文之间多个句子之间的语义一致性. 基于NLI的一致性评价反映生成内容和角色设定之间的一致性. Cao等人[69]和Madotto等人[116]通过训练的NLI模型将生成文本和角色描述的每一个句子进行匹配,NLI模型输出值为1,0,–1分别对应于匹配、独立和矛盾三种类型,所有句子的匹配结果的平均值就是生成文本的角色一致性分值. 基于排序的评价方式也可以用于角色一致性判定. Kim等人[70]针对模型输出的多个生成文本中的角色内容和参考文本进行比较,将与参考文本中角色相关度最高的生成文本在多个生成文本中的排序位置作为MRR分值.

表8 忠实度评价方法

评价类型	评价方法	评价任务
N-gram	BLEU[93], ROUGE[109]	训练数据集中幻觉信息较少的生成任务
	PARENT[95], PARENT-T[117]	数据到文本的生成
基于比率	输入信息覆盖率[40,42,43]	摘要生成、对话回复生成、机器翻译、释义生成
	幻觉比率[39,40,42,43]	
	BERTFaithful[44]	
基于问答	FEQA[96], QuestEval[97]	摘要生成
基于 NLI	RankNLI[98]	摘要生成
基于依赖弧	DAE[101]	摘要生成、释义生成

表9 事实性评价方法

评价类型	评价方法	评价任务
基于比率	F1, Accuracy, Precision, Recall	摘要生成、机器翻译
基于分类	FactCC, FactCCX[102]	摘要生成
	BEGIN[103]	对话回复生成
基于问答	QAGS[104], QUALS[105]	摘要生成
基于 NLI	CI-ToD[106], DialFact[107]	对话回复生成
基于匹配	幻觉实体度量、内容证实度量[108]	开放式文本生成

多样性评价(表12)的基本思想包括三个方面,分别是如何计算生成文本和输入文本之间的语义保留,如何计算具有相似语义的多个生成文本之间的差异,以及如何评价生成文本的风格准确性. 在计算多个生成文本之间差异方面,SELF-BLEU[93]通过采样多个生成文本并计算每两个文本之间的BLEU值,并通过它们的平均值去评价多样性,该值越低说明多样性越好. 而Distinct-n采用计算单个生成文本中不重复的N-gram数量来衡量多样性,Distinct-n值越大,多样性越好. 而Su等人[84]认为句子的熵较大时也能说明较高的多样性,并通过计算生成文本的4-gram熵来给出多样性分值. 生成文本的语义保留评价和文本忠实度的评价思想类似,可以采用以BLEU, METEOR, ROUGE-L和GREEDY MATCH等方法评价文本之间的相似性,这些评价值越高表示语义保留越好. 在计算生成文本的风格准确性时,Cheng等人[82]和Hu等人[85]采用Accuracy值来衡量生成文本是否符合设定的风格类型. 而风格分类器也被用于风格的评价. Zhou等人[24]和Shu等人[79]利用基于TextCNN和LSTM的预训练风格分类器去预测生成文本的风格标签以判断准确性.

流畅度评价(表13)的核心思想是如何计算生成文本中的词语组合概率. 而基于N-gram的词重叠计算就是遵循这种思想的基本评价方法,包括多精度值的加权平均值BLEU[93]、精度与召回率的调和平均值METEOR[110]等. 此外,困惑度(Perplexity,PPL)被认为等价于对数交叉熵,它通过计算生成文本中所有相邻词之间的条件概率来反映生成文本的流畅度. 考虑到文本生成模型通常倾向于生成高概率却缺乏信息量的词汇,Kann等人[118]脱离参考文本的限制,提出的语法对数优势比(Syntax log-odds ratio,SLOR)和WPSLOR利用生成句子的上下文概率和词概率的差值作为流畅度评价,防止了低概率词对整个生成文本评价的负面影响. 基于文本语义计算的方法同样被用于流畅度评价. Zhang等人[119]认为基于N-gram的评价方法可能会使部分语义正确但与参考文献不符的句子得到较低的评分,也无法捕获长文本中的远程依赖关系并惩罚语义上的顺序变化,所以Zhang等人[119]提出的BERTScore模型使用基于BERT的上下文编码来描述生成的句子和参考句子,通过计算两个句子之间的余弦相似度作为流畅度指标.

4.2 人工评价

人工评价的基本思想是多个人类评价者同时对不同模型的生成结果,根据质量特征选择统一的评价方法,最终采用李克特量表打分并计算评价均值. 忠实度注重在生成文本中还原输入内容中的关键信息. Rebuffel等人[39]和Liu等人[42]要求评价者判定生成文本中是否包含幻觉句子并确定输入内容的覆盖度. Rashkin等人[42]还要求评价者分析生成内容是否包含有个人主观的非理性信息. 事实性则侧重于多种细粒度的客观评价. Calderon等人[55]要求评估者判断生成内容的领域正确性、标签类别、语言可接受性、错误率等细粒度

表10　连贯性评价方法

评价类型	评价方法	评价任务
N-gram	BLEU[93]、ROUGE[109]、METEOR[110]	摘要生成、对话回复生成、机器翻译、释义生成
基于比率	F1	对话回复生成、摘要生成、机器翻译
基于语义	Fréchet Distance[18]、LCD[112]	长文本生成
基于实体	实体词的代词共指概率[113,114]	长文本生成、故事续写
基于词链	词链匹配值[115]	机器翻译

表11　一致性评价方法

评价类型	评价方法	评价任务
N-gram	BLEU[93]、ROUGE[107]、METEOR[110]、PPL	摘要生成、对话回复生成
基于比率	F1、Accuracy、Contradict[70]、Macro Avg、Weighted Avg[26]	对话回复生成、故事续写
基于语义	Jaccard	摘要生成、机器翻译
基于 NLI	C score	对话回复生成
基于排序	MRR	对话回复生成

表12　多样性评价方法

评价类型	评价方法	评价目标
N-gram	BLEU[93]、ROUGE[109]、METEOR[110]、PPL、GREEDYMATCH	生成文本的语义保留
	SELF-BLEU[93]、Distinct-n、Entropy[84]	生成文本的差异
基于比率	F1、Accuracy	生成文本的风格准确性
基于分类	风格分类器[24,79]	生成文本的风格准确性

表13　流畅度评价方法

评价类型	评价方法	评价任务
N-gram	BLEU[93]、METEOR[110]、PPL	摘要生成、对话回复生成、机器翻译、长文本生成、故事续写
	SLOR、WPSLOR[118]	文本自动压缩、摘要生成
基于语义	BERTScore[119]	摘要生成

事实性质量. Guan等人[50]则需要评估者评判生成内容在上下文中的因果关系和时间依赖方面是否合理. 连贯性考察生成文本的上下文语义相似性. 如Hua等人[19]认为评价者应当关注于连贯性内容的组织是否自然或是否符合逻辑,涉及的关键词、槽值和文本计划(Plan)对生成内容是否产生连贯性影响. Tan等人[18]要求评估者分别从段级和句子级出发,对长文本连贯性给出评分以计算连贯性内容的百分比. 一致性与连贯性相似却更加注重内在属性. 如Jang等人[20]在问卷中提供了多个生成对话内容和问题,要求评价者进行选择哪一个生成内容回答了问题且符合角色一致性要求. Cao等人[69]则要求评价者关注于生成文本是否和角色描述相匹配. 多样性的人工评价需要同时考虑到生成结果与原始文本的差异以及不同评价者的区别. Cheng等人[82]和Khayrallah等人[83]认为评价者应充分考虑到原始文本的内容,并与其他生成模型对比. Dai等人[81]将不同模型生成的结果打乱后交给评价者以降低偏见,并关注风格、语义保留等多样性质量维度. 流畅度的人工评价聚焦于内容信息表述是否通顺以及语法是否正确. Tan等人[18]关注于生成内容和强基线之间的比较,而非参考文本. Moryossef等人[64]还通过被广泛用于人工语法评价的UPF-FORGe指标,测量生成内容的语法合理性.

此外,ChatGPT[2]和Bard等模型的成功也得益于人类评价者的反馈. 尤其以RLHF (Reinforcement Learning by Human Feedback)技术为代表的人机交互机制,它基于强化学习并构建奖励模型(Reward Model,RM),使模型不仅能够学习到任务核心和人类偏好,而且可以促进Zero-shot问题的解决. 然而,现有的人工评价方法仍然存在3个不足. 首先,人工评价需要准备足够的人力和时间,评价者也需要具有足够的领域知识;其次,即使是具有相似身份的评估者也很难达成一致意见,对于相同的文本通常也会得到波动较大的分数;最后,从质量特征的角度来看,并不是所有质量特征都适合人工评价[32],比如多样性的评价可能会受到评价者的主观感受影响. 所以,如何构建人工评价者组织机制和工作流程,形成统一的行业评价标准应当在未来被重视.

5　未来研究方向

文本内在质量控制一直是自然语言生成领域的重

要研究方向. 不同的质量特征被广泛用于生成模型的设计和评估. 本文将6种质量特征对应于“信、达、雅”三个基本质量需求. Celikyilmaz等人[32]在讨论摘要评估任务时也考虑到了7种质量特征. Li等人[120]在讨论忠实度时也将整体文本生成质量特征分为了4个大类以及多个的小类. 上述研究从不同的视角出发给出了文本生成质量控制的分析和总结,但是不同工作之间所采用的质量特征定义以及归类存在一定的差异,导致这些工作的可比性和对照参考性下降. 特别是在大语言模型的时代,学界和产业界迫切需要系统且统一的文本质量定义,以控制生成模型的设计、训练、测试和评估等工作. 此外,当生成文本需要符合不同领域或工作环境时,还应当充分考虑到以安全性、价值化、功能性、兼容性等外部质量特征要求. 所以如何结合语言学、哲学、传播学、自然语言处理技术,提出一种被广泛接纳的标准化质量特征体系,对大语言模型时代的文本生成技术的未来发展十分重要! 本文根据6种质量特征的模型设计和评价方法的分析,进一步探讨了不同内在质量特征的未来发展方向,并提出了一系列研究内容,希望能够为研究者提供一定的帮助.

5.1 忠实度

文本忠实度的首要挑战就是对抗生成文本中残存的“幻觉”信息,所以需要研究:如何构造系统性的幻觉体系以实现细粒度的幻觉分类;如何定位生成文本中的幻觉位置和跨度,尤其当生成文本中同时存在多个不同类型的幻觉内容;如何提高幻觉检测的可解释性;如何针对幻觉类型,设计文本幻觉纠正模型. 从提升忠实度的当前难点出发,需要研究:如何提升数值类型的忠实度;如何基于不同幻觉类型评价生成文本忠实度.

5.2 事实性

从事实性依赖的“证据”角度出发,需要研究:如何构造具有不同事实置信度的数据集,通过不同事实置信度的数据样本增强模型对“绝对可信、相对可信、可能可信、绝不可信的证据”的理解;如何设计结构化证据和非结构化证据之间的转换模型,以控制事实性生成过程;当证据和训练数据之间存在矛盾,或多个证据之间存在矛盾时,如何选取或改造证据信息. 从事实性文本生成的控制角度看,需要研究:如何融合多种模态的知识以促进基于大语言模型的事实性控制;如何增强反事实生成的可解释性;如何从证据的角度出发,结合大语言模型去评价生成文本的事实性.

5.3 连贯性

大语言模型的出现增强了生成内容的连贯性,也提出了更高的要求. 未来研究包括:如何基于KG增强生成文本的语义连贯性;如何构造提示信息,以控制生成内容中更细粒度的逻辑连贯性;如何利用大语言模型,评价生成文本的语义和逻辑连贯性.

5.4 一致性

在角色一致性方面,需要研究:如何基于大语言模型和知识图谱去增强角色的属性信息,并保持角色属性信息的自洽;如何检测带有多属性角色信息中潜在的矛盾;如何将复杂的角色信息应用于生成控制,并使大语言模型关注到角色属性的细节;如何评价构造的角色质量. 在情感一致性方面,需要研究:如何使生成内容中的情感表达更加符合语境以增强同理心;如何实现情感一致性检测的可解释性,以及文本的情感修正算法.

5.5 多样性

以ChatGPT为代表的一系列大语言模型为多样化的文本生成提供了优秀的平台,但在特定分支领域的多样性生成以及模仿特定人物的语言风格时,仍然需要深入研究基于中小规模生成模型的多样性以提高准确性和适应性. 具体还应当研究:如何构造核心语义的提取机制并提高风格解缠过程的可解释性;如何控制生成文本的风格或样式,以避免风格转换不完全和风格混乱的问题;如何实现多种类型风格的组合控制,以实现更为精细的多样性内容生成;如何结合多模态信息去控制文本多样性生成;以及如何通过文本多样性去促进多模态内容的合成.

5.6 流畅度

大语言模型的出现使生成文本具备优秀的可读性和上下文逻辑,但也存在训练开销、部署成本和价格方面的约束. 所以未来需要研究包括:如何提升中小语言模型的生成文本的流畅度,以接近大语言模型的生成能力,并保持在特定领域任务下的信息量;如何实现少数国家或民族语言的文本流畅度,尤其是缺乏训练数据集的模型训练和微调;多语言混合的生成任务中,如何借助大语言模型语义理解和意图对齐能力,在保证生成文本的流畅度的同时提升信息量.

6 结论

基于文本内在质量约束的内容生成和评价是AIGC在文本领域中的重要环节. 本文从“信、达、雅”的质量需求出发,将内在质量约束分解为包括忠实度、事实性、连贯性、一致性、多样性、流畅度在内的6种质量特征. 本文提出的研究问题1聚焦不同质量特征下的文本生成模型和算法设计:本文基于不同质量特征的定义,对比并分析了相关模型的设计思想和性能. 本文提出的研究问题2聚焦生成文本的内在质量评估:本文分别以自动与人工方法对6项质量特征的评价机制进行了总结. 希望通过上述工作为设计出更高质量的生成模型提供参考.

参考文献

[1] OUYANG, WU J, JIANG X, et al. Training language models to follow instructions with human feedback [EB/OL]. (2022-03-04) [2023-09-01]. https://arxiv.org/abs/2203.02155.

[2] SCHULMAN J, ZOPH B, KIM C, et al. ChatGPT: Optimizing language models for dialogue[R/OL]. (2022-11-30) [2023-09-01]. https://openai.com/blog/chatgpt.

[3] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 1877-1901.

[4] LEWIS M, LIU Y H, 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. Stroudsburg: Association for Computational Linguistics, 2020: 7871-80.

[5] TURING A M. Computing machinery and intelligence [M]//EPSTEIN R, ROBERTS G, BEBER G. Parsing the Turing Test. Dordrecht: Springer, 2009: 23-65.

[6] ZHANG M. An inquiry into Yan fu's translation theory of faithfulness, expressiveness, and elegance: The beginning of China's modern translation theory[J]. Trans-Humanities Journal, 2013, 6(3): 179-196.

[7] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[EB/OL]. (2019-10-23) [2023-09-01]. https://arxiv.org/abs/1910.10683.

[8] BUBECK S, CHANDRASEKARAN V, ELDAN R, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4[EB/OL]. (2023-03-22) [2023-09-01]. http://arxiv.org/abs/2303.12712.pdf.

[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017-06-12) [2023-09-01]. http://arxiv.org/abs/1706.03762.pdf.

[10] DEVLIN J, CHANG M-W, LEE K, 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: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019: 4171-4186.

[11] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[EB/OL]. (2018-10-11) [2023-09-01]. https://cdn. openai. com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[12] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[EB/OL]. (2019-02-14) [2023-09-01]. https://d4mucfpksywv. cloudfront. net/better-language-models/language-models.pdf.

[13] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: A robustly optimized BERT pretraining approach[EB/OL]. (2019-07-26)[2023-09-01]. https://arxiv.org/abs/1907.11692.

[14] JOSHI M, CHEN D Q, LIU Y H, et al. SpanBERT: Improving pre-training by representing and predicting spans [J]. Transactions of the Association for Computational Linguistics, 2020, 8: 64-77.

[15] 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 (EMNLP-IJCNLP). Stroudsburg: Association for Computational Linguistics, 2019: 3980-3990.

[16] DAI Z H, YANG Z L, YANG Y M, et al. Transformer-XL: Attentive language models beyond a fixed-length context[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 2978-2988.

[17] YANG Z L, DAI Z H, YANG Y M, et al. XLNet: Generalized autoregressive pretraining for language understanding[C]//NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems. New York: ACM, 2019: 5753-5763.

[18] TAN B W, YANG Z C, AI-SHEDIVAT M, et al. Progressive generation of long text with pretrained language models[EB/OL]. (2020-06-28) [2023-09-01]. https://arxiv.org/abs/2006.15720.

[19] HUA X Y, WANG L. PAIR: Planning and iterative refinement in pre-trained transformers for long text generation[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Kerrville: Association for Computational Linguistics, 2020: 781-793.

[20] JANG Y, LIM J, HUR Y, et al. Call for customized conversation: Customized conversation grounding persona and knowledge[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(10): 10803-10812.

[21] CHEN Z, XU W D, WANG B T, et al. A blockchain-based preserving and sharing system for medical data privacy[J]. Future Generation Computer Systems, 2021, 124 (C): 338-350.

[22] LE T, PARK N, LEE D. SHIELD: defending textual neu-

ral networks against multiple black-box adversarial attacks with stochastic multi-expert patcher[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2022: 6661-6674.

[23] YI X Y, LIU Z H, LI W H, et al. Text style transfer via learning style instance supported latent space[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2020: 3801-3807.

[24] ZHOU C L, CHEN L Y, LIU J C, et al. Exploring contextual word-level style relevance for unsupervised style transfer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 7135-7144.

[25] MO L Z, WEI J L, HUANG Q B, et al. Incorporating sentimental trend into gated mechanism based transformer network for story ending generation[J]. Neurocomputing, 2021, 453(C): 453-464.

[26] WEN Z Y, CAO J N, YANG R S, et al. Automatically select emotion for response via personality-affected emotion transition[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Stroudsburg: Association for Computational Linguistics, 2021: 5010-5020.

[27] YIN X J, WAN X J. How do Seq2Seq models perform on end-to-end data-to-text generation? [C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2022: 7701-7710.

[28] SHU C, ZHANG Y S, DONG X Y, et al. Logic-consistency text generation from semantic parses[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Stroudsburg: Association for Computational Linguistics, 2021: 4414-4426.

[29] DER LEE CHRIS V, ALBERT G, EMIEL V M, et al. Human evaluation of automatically generated text: Current trends and best practice guidelines[J]. Computer Speech & Language, 2021, 67: 101151.

[30] IQBAL T, S J J O K S U-C QURESHI, SCIENCES I. The survey: Text generation models in deep learning[J]. Journal of King Saud University - Computer and Information Sciences, 2020, 34(6): 2515-2528.

[31] LI J Y, TANG T Y, ZHAO W X, et al. Pretrained language model for text generation: A survey[C]//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2021: 4492-4499.

[32] CELIKYILMAZ A, CLARK E, GAO J F. Evaluation of text generation: A survey[EB/OL]. (2020-06-26) [2023-09-01]. http://arxiv.org/abs/2006.14799.pdf.

[33] JIN D, JIN Z J, HU Z T, et al. Deep learning for text style transfer: A survey[J]. Comput Linguistics, 2022, 48(1): 155-205.

[34] ZHOU J N, BHAT S. Paraphrase generation: A survey of the state of the art[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 5075-5086.

[35] ZHAO H, PHUNG D, HUYNH V, et al. Topic modelling meets deep neural networks: A survey[C]//Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2021: 4713-4720.

[36] JI Z W, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2022, 55(12): 248.

[37] TANG C, GUERIN F, LIN C H. Recent advances in neural text generation: A task-agnostic survey[EB/OL]. (2022-03-06)[2023-09-01]. http://arxiv.org/abs/2203.03047.pdf.

[38] HALLIDAY M A K, MATTHIESSEN C M I M. Halliday's Introduction to Functional Grammar[M]. London: Routledge, 2013.

[39] REBUFFEL C, ROBERTI M, SOULIER L, et al. Controlling hallucinations at word level in data-to-text generation[J]. Data Mining and Knowledge Discovery, 2022, 36 (1): 318-354.

[40] RASHKIN H, REITTER D, TOMAR G S, et al. Increasing faithfulness in knowledge-grounded dialogue with controllable features[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2021: 704-718.

[41] LI H R, ZHU J N, ZHANG J 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. Stroudsburg: Association for Computational Linguistics, 2018: 1430-1441.

[42] LIU T Y, ZHENG X, CHANG B B, et al. Towards faithfulness in open domain table-to-text generation from an entity-centric view[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(15): 13415-13423.

[43] ZHANG J Q, ZHAO Y, SALEH M, et al. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization[C]//Proceedings of the 37th International Conference on Machine Learning. New York: ACM, 2020: 11328-11339.

[44] ARALIKATTE R, NARAYAN S, MAYNEZ J, et al. Focus attention: Promoting faithfulness and diversity in summarization[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2021: 6078-6095.

[45] CAO S Y, WANG L. CLIFF: contrastive learning for improving faithfulness and factuality in abstractive summarization[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 6633-6649.

[46] LEE S, LEE D B, HWANG S J. Contrastive learning with adversarial perturbations for conditional text generation[EB/OL]. (2020-12-14)[2023-09-01]. http://arxiv.org/abs/2012.07280.pdf.

[47] ZHU C G, XU Y C, REN X, et al. Knowledge-augmented methods for natural language processing[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts. Stroudsburg: Association for Computational Linguistics, 2022: 12-20.

[48] ZHOU H, YOUNG T, HUANG M L, et al. Commonsense knowledge aware conversation generation with graph attention[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. New York: ACM, 2018: 4623-4629.

[49] TAI Y, YANG J, LIU X M, et al. MemNet: A persistent memory network for image restoration[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 4549-4557.

[50] GUAN J, HUANG F, ZHAO Z H, et al. A knowledge-enhanced pretraining model for commonsense story generation[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 93-108.

[51] YU C L, ZHANG H M, SONG Y Q, et al. CoCoLM: complex commonsense enhanced language model with discourse relations[C]//Findings of the Association for Computational Linguistics: ACL 2022. Stroudsburg: Association for Computational Linguistics, 2022: 1175-1187.

[52] MADAAN N, PADHI I, PANWAR N, et al. Generate your counterfactuals: Towards controlled counterfactual generation for text[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(15): 13516-13524.

[53] PARANJAPE B, LAMM M, TENNEY I. Retrieval-guided counterfactual generation for QA[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2022: 1670-1686.

[54] BARTOLO M, THRUSH T, JIA R, et al. Improving question answering model robustness with synthetic adversarial data generation[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 8830-8848.

[55] CALDERON N, BEN-DAVID E, FEDER A, et al. DoCoGen: domain counterfactual generation for low resource domain adaptation[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2022: 7727-7746.

[56] ZHANG H N, LAN Y Y, PANG L, et al. Modeling topical relevance for multi-turn dialogue generation[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2020: 3737-3743.

[57] SERBAN I, SORDONI A, BENGIO Y, et al. Building end-to-end dialogue systems using generative hierarchical neural network models[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2016, 30(1): 3776-3784.

[58] HE G X, GAO Z, JIANG Z R, et al. Think beyond the word: Understanding the implied textual meaning by digesting context, local, and noise[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2020: 2297-2306.

[59] QIN L H, GALLEY M, BROCKETT C, et al. Conversing by reading: Contentful neural conversation with on-demand machine reading[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 5427-5436.

[60] LIU X D, SHEN Y L, DUH K, et al. Stochastic answer networks for machine reading comprehension[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2018: 1694-1704.

[61] TIAN Z L, BI W, LEE D, et al. Response-anticipated memory for on-demand knowledge integration in response generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 650-659.

[62] XU J, LEI Z Y, WANG H F, et al. Discovering dialog structure graph for coherent dialog generation[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2021: 1726-1739.

[63] PENG B L, LI C Y, LI J C, et al. SOLOIST: Few-shot task-oriented dialog with A single pre-trained auto-regressive model[EB/OL]. (2020-05-11) [2023-09-01]. http://arxiv.org/abs/2005.05298.pdf.

[64] MORYOSSEF A, GOLDBERG Y, DAGAN I. Step-by-step: Separating planning from realization in neural data-to-text generation[EB/OL]. (2019-08-06) [2023-09-01]. http://arxiv.org/abs/1904.03396.pdf.

[65] HUA X Y, HU Z, WANG L. Argument generation with retrieval, planning, and realization[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 2661-1672.

[66] MARCHENKO O O, RADYVONENKO O S, IGNATOVA T S, et al. Improving text generation through introducing coherence metrics[J]. Cybernetics and Systems Analysis, 2020, 56(1): 13-21.

[67] WEI J, WANG X Z, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[EB/OL]. (2022-01-28)[2023-09-01]. https://arxiv.org/abs/2201.11903.

[68] YAO S Y, YU D, ZHAO J, et al. Tree of thoughts: Deliberate problem solving with large language models[EB/OL]. (2023-05-17) [2023-09-01]. http://arxiv. org/abs/2305.10601.pdf.

[69] CAO Y, BI W, FANG M, et al. A model-agnostic data manipulation method for persona-based dialogue generation[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2022: 7984-8002.

[70] KIM M, KWAK B W, KIM Y, et al. Dual task framework for improving persona-grounded dialogue dataset [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(10): 10912-10920.

[71] FU T C, ZHAO X L, TAO C Y, et al. There are a thousand hamlets in a thousand People´s eyes: Enhancing knowledge-grounded dialogue with personal memory[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2022: 3901-3913.

[72] YANG R Z, CHEN J X, NARASIMHAN K. Improving dialog systems for negotiation with personality modeling [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2021: 681-693.

[73] SONG H Y, WANG Y, ZHANG K Y, et al. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2021: 167-177.

[74] GU J C, TAO C Y, LING Z H, et al. MPC-BERT: A pre-trained language model for multi-party conversation understanding[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2021: 3682-3692.

[75] PARK J S, O´BRIEN J, CAI C J, et al. Generative agents: Interactive simulacra of human behavior[C]//Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. New York: ACM, 2023: 1-22.

[76] WANG Z L, CHIU Y Y, CHIU Y C. Humanoid agents: Platform for simulating human-like generative agents[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Stroudsburg: Association for Computational Linguistics, 2023: 167-176.

[77] SCLAR M, KUMAR S, WEST P, et al. Minding language models´ (lack of) theory of mind: A plug-and-play multi-character belief tracker[C]//Proceedings of the 61st

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2023: 13960-13980.

[78] LEE J. Stable style transformer: Delete and generate approach with encoder-decoder for text style transfer[C]// Proceedings of the 13th International Conference on Natural Language Generation. Stroudsburg: Association for Computational Linguistics, 2020: 195-204.

[79] SHU L, PAPANGELIS A, WANG Y C, et al. Controllable text generation with focused variation[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: Association for Computational Linguistics, 2020: 3805-3817.

[80] SOHN K, YAN X C, LEE H. Learning structured output representation using deep conditional generative models [J]. Advances in Neural Information Processing Systems, 2015, 2015-January: 3483-3491.

[81] DAI N, LIANG J Z, QIU X P, et al. Style transformer: Unpaired text style transfer without disentangled latent representation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 5997-6007.

[82] CHENG Y, GAN Z, ZHANG Y Z, et al. Contextual text style transfer[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: Association for Computational Linguistics, 2020: 2915-2924.

[83] KHAYRALLAH H, SEDOC J. SMRT chatbots: Improving non-task-oriented dialog with simulated multiple reference training[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: Association for Computational Linguistics, 2020: 4489-4505.

[84] SU H, SHEN X Y, ZHAO S Q, et al. Diversifying dialogue generation with non-conversational text[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 7087-7097.

[85] HU Z T, YANG Z C, LIANG X D, et al. Toward controlled generation of text[C]//Proceedings of the 34th International Conference on Machine Learning - Volume 70. New York: ACM, 2017: 1587-1596.

[86] KINGMA D P, WELLING M. Auto-encoding variational Bayes[EB/OL]. (2013-12-20) [2023-09-01]. http://arxiv.org/abs/1312.6114.pdf.

[87] LACHAUX M A, JOULIN A, LAMPLE G. Target conditioning for one-to-many generation[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: Association for Computational Linguistics, 2020: 2853-2862.

[88] CUI Z, LI Y R, ZHANG J Y, et al. Focus-constrained attention mechanism for CVAE-based response generation [C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: Association for Computational Linguistics, 2020: 2021-2030.

[89] WANG Z, RAO S W, ZHANG J, et al. Diversify question generation with continuous content selectors and question type modeling[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: Association for Computational Linguistics, 2020: 2134-2143.

[90] SUN H, LIN Z R, ZHENG C J, et al. PsyQA: A Chinese dataset for generating long counseling text for mental health support[C]//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Stroudsburg: Association for Computational Linguistics, 2021: 1489-1503.

[91] XU J, SZLAM A, WESTON J. Beyond goldfish memory: Long-term open-domain conversation[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2022: 5180-5197.

[92] GU N L, ASH E, HAHNLOSER R. MemSum: extractive summarization of long documents using multi-step episodic Markov decision processes[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2022: 6507-6522.

[93] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: A method for automatic evaluation of machine translation [C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. New York: ACM, 2002: 311-318.

[94] FABBRI A, WU C S, LIU W H, et al. QAFactEval: improved QA-based factual consistency evaluation for summarization[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2022: 2587-2601.

[95] DHINGRA B, FARUQUI M, PARIKH A, et al. Handling divergent reference texts when evaluating table-to-text generation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019:

4884-4895.
[96] DURMUS E, HE H, DIAB M. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 5055-5070.
[97] SCIALOM T, DRAY P A, LAMPRIER S, et al. QuestEval: summarization asks for fact-based evaluation[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 6594-6604.
[98] FALKE T, RIBEIRO L F R, UTAMA P A, et al. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 2214-2220.
[99] MAYNEZ J, NARAYAN S, BOHNET B, et al. On faithfulness and factuality in abstractive summarization[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 1906-1919.
[100] NIE Y X, WILLIAMSON M, BANSAL M, et al. I like fish, especially dolphins: Addressing Contradictions in Dialogue Modeling[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2021: 1699-1713.
[101] GOYAL T, DURRETT G. Evaluating factuality in generation with dependency-level entailment[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: Association for Computational Linguistics, 2020: 3592-3603.
[102] KRYSCINSKI W, MCCANN B, XIONG C M, et al. Evaluating the factual consistency of abstractive text summarization[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2020: 9332-9346.
[103] DZIRI N, RASHKIN H, LINZEN T, et al. Evaluating groundedness in dialogue systems: The BEGIN benchmark[EB/OL]. (2021-04-30) [2023-09-01]. http://arxiv.org/abs/2105.00071.pdf.
[104] 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. Stroudsburg: Association for Computational Linguistics, 2020: 5008-5020.
[105] NAN F, NOGUEIRA DOS SANTOS C, ZHU H H, et al. Improving factual consistency of abstractive summarization via question answering[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2021: 6881-6894.
[106] QIN L B, XIE T B, HUANG S J, et al. Don´t be contradicted with anything! CI-ToD: Towards benchmarking consistency for task-oriented dialogue system[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2021: 2357-2367.
[107] GUPTA P, WU C S, LIU W H, et al. DialFact: A benchmark for fact-checking in dialogue[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2022: 3785-3801.
[108] LEE N, PING W, XU P, et al. Factuality enhanced language models for open-ended text generation[EB/OL]. (2022-06-09)[2023-09-01]. http://arxiv.org/abs/2206.04624.pdf.
[109] LIN C Y. ROUGE: A package for automatic evaluation of summaries[C]//Proceedings of the Text Summarization Branches Out. Stroudsburg: Association for Computational Linguistics, 2004: 74-81.
[110] BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Stroudsburg: Association for Computational Linguistics, 2005:65-72.
[111] KARAKANTA A, GAIDO M, NEGRI M, et al. Between flexibility and consistency: Joint generation of captions and subtitles[C]//Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021). Stroudsburg: Association for Computational Linguistics, 2021: 215-225.
[112] XU P, SAGHIR H, KANG J S, et al. A cross-domain transferable neural coherence model[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 678-687.
[113] ELSNER M, CHARNIAK E. Coreference-inspired coherence modeling[C]//Proceedings of the Annual Meet-

ing of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2008: 41-44.

[114] ROEMMELE M, GORDON A, SWANSON R. Evaluating story generation systems using automated linguistic analyses[C]//Proceedings of the SIGKDD 2017 Workshop on Machine Learning for Creativity. New York: ACM, 2017: 13-17.

[115] GONG Z X, ZHANG M, ZHOU G D. Document-level machine translation evaluation with gist consistency and text cohesion[C]//Proceedings of the Second Workshop on Discourse in Machine Translation. Stroudsburg: Association for Computational Linguistics, 2015: 33-40.

[116] MADOTTO A, LIN Z J, WU C S, et al. Personalizing dialogue agents via meta-learning[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 5454-5459.

[117] WANG Z Y, WANG X Y, AN B, et al. Towards faithful neural table-to-text generation with content-matching constraints[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 1072-1086.

[118] KANN K, ROTHE S, FILIPPOVA K. Sentence-level fluency evaluation: References help, but can be spared! [C]//Proceedings of the 22nd Conference on Computational Natural Language Learning. Stroudsburg: Association for Computational Linguistics, 2018: 313-323.

[119] ZHANG T, KISHORE V, WU F, et al. Bertscore: Evaluating text generation with BERT[C]//Proceedings of the 8th International Conference on Learning Representations. Appleton: ICLR, 2020: 1-43.

[120] LI W, WU W H, CHEN M Y, et al. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods[EB/OL]. (2022-03-10)[2023-09-01]. http://arxiv.org/abs/2203.05227.pdf.

作者简介

兰玉乾　男,1990年4月出生于陕西省西安市.2017年毕业于西北工业大学软件学院软件工程专业.现为西安交通大学软件学院博士研究生.主要研究方向为自然语言处理、文本生成、对话系统和社会智能等.

E-mail: Yuqian_Lan_xjtu@stu.xjtu.edu.cn

饶　元　男,1973年2月出生于湖北省武汉市.现为西安交通大学电信学部软件学院教授、博士生导师.陕西省人工智能联合(重点)实验室秘书长、副主任,西安市社会智能与复杂数据处理重点实验室主任.获得军委军事科学与技术进步奖、中国发明协会、陕西省教育厅高校优秀成果、王宽诚育才奖等20余项.国内外发表学术论文70余篇.申请软件著作权30余项,专利24项.主要研究方向为文本数据挖掘、自然语言处理、机器学习以及社会网络分析等.

E-mail: yuanrao@163.com

李冠呈　男,1991年9月出生于山东省日照市.2018年毕业于北京理工大学自动化学院.现为中国长峰机电技术研究设计院工程师.主要研究方向为体系设计.

E-mail: 414773396@qq.com

孙　菱　女,1997年出生于四川省广元市.2019年毕业于中央民族大学软件工程专业.现于西安交通大学软件学院攻读博士学位.主要研究方向为复杂网络和虚假信息传播.

E-mail: sunling@stu.xjtu.edu.cn

夏昺灿　男,1997年12月出生于河南省灵宝市.2019年本科毕业于西北农林科技大学信息工程学院计算机科学与技术专业.2023年硕士毕业于西安交通大学软件学院软件工程专业.现为西安交通大学博士研究生.主要研究方向为图神经网络.

E-mail: bingcan92@xjtu.stu.edu.cn

辛婷婷　女,1995年8月出生于山西省运城市.2021年毕业于西安工程大学计算机科学与技术学院.现为西安交通大学博士研究生.主要研究方向为社交网络中信息传播策略.

E-mail: xinting828@gmail.com