

智能文本生成：进展与挑战

万小军

北京大学王选计算机研究所, 北京 100080

摘要

智能文本生成是人工智能与自然语言处理领域的前沿研究方向, 也是AI生成内容(AIGC)的关键技术支持, 能够大幅提升文本内容的生成效率, 近年受到学术界和产业界的高度关注, 在媒体出版、电子商务等多个行业与场景均已得到应用。对智能文本生成的应用现状与主要方式进行系统概述, 重点介绍基于深度学习的智能文本生成技术, 同时阐述现有技术面临的挑战。

关键词

智能文本生成; 自然语言生成; 机器写作; 自然语言处理

中图分类号: TP39

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2023014

Intelligent text generation: recent advances and challenges

WAN Xiaojun

Wangxuan Institute of Computer Technology, Peking University, Beijing 100080, China

Abstract

Intelligent text generation is one of the advanced research directions in the fields of artificial intelligence and natural language processing. It is one of key technologies for making AIGC successful, and it has been highly concerned by both academia and industry in recent years. The technology has been deployed and used in many application areas including media publishing and e-commerce, and it can much improve the efficiency of text content production. The systematic overview of the applications of intelligent text generation and the mainstream ways of text generation were given. And the recent deep learning techniques for text generation were introduced. Lastly, the challenges faced by neural text generation were summarized.

Key words

intelligent text generation, natural language generation, machine writing, natural language processing

0 引言

随着深度学习技术的发展和应⤵, AI生成内容 (AI-generated content, AIGC) 的能力得到显著提升。文本、图像、视频、音频等内容的自动生成已成为人工智能领域的热门研究方向, 同时也备受产业界的关注, 在各行各业均具有重要的应用价值。

语言文字是人类最重要的交流工具和思想文化载体, 实现语言智能是人工智能走向成熟的重要标志, 智能文本生成则是语言智能的关键支撑技术。智能文本生成又称自然语言生成或机器写作, 其目的是根据给定的输入数据 (如报表数据、视觉信息、意义表示、文本素材等) 自动生成高质量的不同类型的自然语言语句或篇章 (如标题、摘要、新闻、故事、诗歌、评论、广告等)^①。对于自动生成的文本, 一般要求可读性好, 同时内容准确可靠。需要指出的是, 由于语言表达的多样性, 文本生成的结果通常是不唯一的, 特别是对于开放式文本生成任务 (如文本复述、故事生成等) 而言, 同一输入生成的输出可以有成百上千种可能。举例来说, 对于文本复述任务而言, 针对输入语句“梅西是足坛的GOAT”, 该语句可以被改写为“梅西是足坛历史最佳”, 或者“足坛历史第一人是梅西”, 或者“梅西的成就超越了足坛所有其他球员”, 等等。

文本生成的困难之处主要有两点。一是搜索空间巨大。文本自动生成可以看作一个搜索问题, 对于长度为 L 的文本 (即文本包含 L 个词), 假设词表大小为 W (即共有 W 个词语), 考虑到文本的每个位置都可以从 W 个词语中任选一个词语进行填充, 那么总共有 W^L 种可能的组合。 W 通常很大 (一般为几万到几十万的规模), 因

此这个搜索空间巨大, 这给文本生成算法带来了严重挑战。二是文本质量难以客观评价。尽管我们能够主观判断一个语句或一篇文章的优劣, 但是很难找到一个自动评价指标对文本质量进行客观、准确的评估。目前业界使用的基于 N 元词重叠程度的评价指标 (如BLEU、ROUGE等) 与模型驱动的评价指标 (如BERTScore、BARTScore等) 均难以可靠地评价文本质量 (特别是开放式文本生成任务及长文本生成任务), 导致文本生成任务的优化目标难以准确定义和形式化, 现有的基于最大似然的优化目标与文本生成的总体质量目标之间存在相当大的差异, 这也是现有文本生成模型一个不可忽视的缺陷。

本文后续章节将对智能文本生成的应用现状、主要方式进行概述, 重点介绍基于深度学习的智能文本生成技术, 同时指出智能文本生成技术面临的挑战, 最后进行总结与展望。

1 智能文本生成的应用现状

近几年智能文本生成技术有了突破性发展, 各类智能写作需求广泛崛起, 智能文本生成应用呈现出行业广、场景多、需求大等特点。智能文本生成已经在媒体出版、电子商务、人机交互、电子政务、智慧教育、智慧医疗、智慧司法等多个行业和领域得到了落地应用。国内外数十家单位和企业 (如OpenAI、ARRIA、AutomatedInsights、NarrativeScience、Google、Microsoft、阿里、百度、腾讯、京东等) 将文本生成能力作为核心竞争力之一, 已相继推出各类文本内容生成工具与服务, 能够自动化生产或辅助人工生产各类文本内容 (包括新闻、财报、

① 考虑到绝大部分文本生成任务与应用属于有条件文本生成范畴, 本文只关注有条件文本生成, 而忽略无条件文本生成 (根据随机噪声生成文本)。

天气预报、文摘、会议纪要、综述、公文、产品说明、广告文案、对话回复、评论等），大幅提升了内容生产效率和覆盖率。特别是在OpenAI推出ChatGPT之后，业界对文本生成模型的强大能力有了更深刻的认识，因而对智能写作技术的应用有了更大的想象空间，例如辅助写论文、发言稿、小说等。下面以媒体出版和电子商务两个典型领域为例具体介绍智能文本生成技术的应用情况。

媒体出版是智能文本生成最重要和最典型的应用领域。首先，许多传统媒体单位（如报社）逐步采用智能文本生成技术进行新闻稿件的自动或辅助创作，主要面向结构和表达相对简单的天气预报、赛事简讯、财经简讯等新闻类型，同时为新闻自动生成标题、摘要等，方便读者快速浏览新闻内容。例如，美联社自2014年7月开始采用新闻写作软件自动撰写新闻稿件报道公司业绩，大大降低了记者的工作量。2019年，笔者与科学网进行了面向科技新闻类型文本生成的全新尝试与合作，研制推出了小柯写作机器人。小柯是一个跨语言科技新闻生成系统，能够基于机器翻译和文本摘要技术自动根据英文学术论文生成简短的中文科技新闻。目前小柯机器人已覆盖物理学、地球科学、生命科学、化学科学4个学科，撰写科技新闻1万多篇，总阅读量超过1 000万次。考虑到科学网的专业性和严肃性，所有自动生成的稿件在发布前均需编辑进行人工审核。其次，互联网上出现越来越多的新媒体平台，为广大网民提供新闻资讯服务，这类平台通常采用信息推荐技术将不同话题的新闻推荐给不同用户，从而增强用户对平台的黏性和忠诚度，提高平台流量。由于新媒体平台的用户数量庞大，用户感兴趣的话题也不尽相同，因此对新闻稿件的需求量也会很大，采用人工编辑撰写新闻的传统方式已无法满足

用户需求，亟须采用智能文本生成技术批量生产大量新闻稿件并进行分发。可以看到，目前很多新媒体、自媒体平台上均有大量新闻由算法生成，覆盖民生、娱乐、体育、财经、社会等各领域。考虑到新媒体用户对新闻质量的容忍度较高，自动生成的新闻通常不需要人工审核即可发布。例如，2016年里约热内卢奥运会期间，笔者与字节跳动合作研制了小明写作机器人，该款机器人主要撰写体育赛事新闻^②，是业界首款既能创作简讯又能创作长篇报道的写作机器人。其中体育赛事简讯基于体育赛事数据自动生成，长篇报道则基于体育直播文字进行智能筛选与组合而成。小明机器人撰写的体育新闻自动发布到今日头条平台，服务亿万用户。此外，部分出版社尝试利用智能文本生成技术进行书籍的自动编写，目前主要用于编写与科技进展介绍相关的书籍。例如，2019年施普林格出版社与德国法兰克福大学合作，开发算法自动编写了一本介绍锂离子电池最新研究进展的书籍，该算法主要对该学科大量现有的学术论文进行聚类 and 摘要自动生成。智能文本生成技术能够驱动媒体出版行业的流程变革，节约人力和成本，而随着智能文本生成技术的快速发展，写作机器人将有能力撰写结构更复杂、更有深度的新闻稿件，进一步扩大应用范围。

电子商务是智能文本生成另一个应用比较成熟的领域。首先，主流电商网站（如京东、淘宝）会定期上线大量的全新商品，这些商品通常需要对应的标题及描述，从而方便用户浏览和购买。考虑到商品数量巨大，人工撰写的成本会很高，而且效率低下，因此主流电商网站通常会采用智能文本生成技术基于商品的基本参数进行商品标题和描述的自动撰写。目前电商网站上的商品标题和描述大多是通过算法自动生成的，一般具有较高的质量。其次，电商

② 小明机器人的写稿类型后续进一步拓展到财经新闻，并且覆盖 10 多种语言。

网站为了应对用户在购物、售后等方面的交流需求，通常会部署智能客服系统回复用户的问题。智能客服系统需要精准理解用户意图，并利用文本生成技术生成相应的回复文本。同时，部分电商网站会基于对话摘要技术对客服与用户的对话交流进行总结，生成简短的摘要。最后，为了对商品和服务进行推广和营销，很多商家会采用智能文本生成技术自动为商品生成广告和营销类文本，然后发布到众多媒体平台，吸引用户的注意，从而提升商品销量。可以看到，智能文本生成技术已渗透、应用到电子商务的众多流程环节，成为电子商务发展的助推器。

在智能文本生成技术的应用过程中，需要甄别真需求，避免伪需求。真需求即高频率的写作需求，伪需求则是低频率的写作需求。某些场景下年均只需要撰写少数几篇稿件，这种写作需求并不需要采用智能文本生成技术来应对，因为智能写作系统带来的效益会低于开发智能写作系统的投入。在甄别出真实的智能写作需求之后，智能写作系统的开发方通常需要与应用方进行深入沟通和合作，因为智能写作系统的成功离不开应用方的数据和经验。此外，目前智能写作的技术水平与若干智能写作应用需求之间存在较大的差距，也就是说，目前的智能文本生成技术无法满足若干复杂的智能写作需求，例如，自动撰写爆款网络小说、自动撰写长篇科普文章等。

2 智能文本生成的主要方式

从文本生成输入和输出之间的关系出发，本文将智能文本生成的主要实现方式归为以下3类^③。

其一，基于文本扩写的智能文本生

成。文本扩写即文本从无到有，或者从少到多，该方式以非文本信息或少量文字信息为输入进行文本生成。基于数据的文本生成(Data2Text)、基于AMR的文本生成(AMR2Text)、基于视觉的文本生成(Visual2Text，主要包括图像描述生成和视频描述生成)、故事/作文生成(基于标题或关键词的篇章生成)等属于文本扩写方式。该方式的特点是输出文本中的大量或全部文字需要由算法模型进行补充或创作。

其二，基于文本缩写的智能文本生成。文本缩写即文本从多变少，该方式对输入的文本内容进行浓缩和提炼，并输出相对较短的文本，主要包括文本摘要生成、标题生成、综述生成等任务。该方式的特点是与输出文本相比，输入文本通常包含更多的信息，因此需要算法模型进行信息的重要性评估和筛选，使输出文本尽可能包含输入文本中的重要内容，同时舍弃边缘内容。

其三，基于文本改写的智能文本生成。文本改写即对文本进行改变以达到特定目标，输出文本与输入文本的长度相差不会很大。属于该方式的文本生成任务主要包括文本复述、文本简化、文本风格迁移等。该方式的特点为输出文本与输入文本通常可以进行词汇短语级别的对齐，可看作单语言机器翻译问题。

图1总结了3类主要智能文本生成方式及涵盖的典型文本生成任务。上述3类智能文本生成方式可独立应用于智能写作系统，也可进行组合，实现更加复杂的机器写作目标。例如，可以结合扩写与缩写两种方式进行财经类新闻的自动生成。具体来说，可基于Data2Text技术针对财经类数据进行描述生成，然后基于文本摘要技术总结专家观点，将两者结合在一起，形成同时包含基本财经信息和专家观点的财经新闻，更好地服务读者。

③ 对话生成(包括闲聊型、任务完成型)也属于智能文本生成范畴，但因其自身已形成一个相对独立且活跃的研究领域，因篇幅有限，本文不再列入及阐述。

3 基于深度学习的智能文本生成

3.1 文本生成方法概述

早期的文本生成方法主要基于规则，并且采用流水线框架，将复杂的文本生成任务分解为多个阶段，对每个阶段分别进行设计和求解。其优点是不依赖大规模标注数据，可解释性较好，错误容易诊断和定位；缺点是过于依赖专家知识，领域迁移性差，流水线框架也会带来错误累积。例如，对于Data2Text任务而言，英国阿伯丁大学的Reiter E提出了基于流水线的生成框架，包括信号分析(signal analysis)、数据阐释(data interpretation)、文档规划(document planning)、微规划与实现(microplanning and realisation) 4个阶段^①，并可以进一步将该框架简化为数据分析、文档规划、语句实现3个阶段。其中数据分析的目的为对输入数据进行分析理解，获取高层消息以及消息之间的关系；文档规划的目的为分析决定哪些消息和关系需要在文本中被提及，同时确定文本的整体结构、输出需要提及的消息及文档结构；语句实现的目的是基于选中的消息及结构进行句子规划及实现，要求最终实现的句子具有正确的语法、形态和拼写，同时采用准确的指代表达。上述框架中的每个阶段都可以考虑基于规则实现，对于语句实现可考虑利用文法信息。随着标注数据的积累和增多，业界逐步采用统计方法进行文本生成。考虑到文本生成与机器翻译的相似性（将输入信息翻译/转换为特定自然语言表达），不少研究者尝试复用机器翻译模型来完成文本生成。需要说

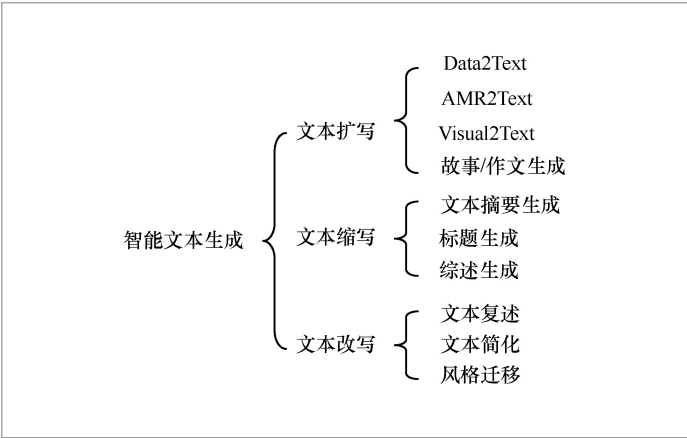


图1 智能文本生成主要方式及典型文本生成任务

明的是，不同文本生成任务具有不同的特点，因此不同文本生成任务的阶段分解、规则设计及适用的统计方法不尽相同，需要有针对性地考虑。

近几年，随着深度学习技术的发展和突破，基于深度学习的文本生成已成为最主流的技术路线，每年在领域重要国际会议(包括ACL、EMNLP、NAACL、EACL、AAACL、AAAI、IJCAI、INLG 等)均有大量相关学术论文发表，推动文本生成任务的技术创新和性能提升。深度学习技术可用于流水线框架的各个阶段，但更主流的做法通常是将文本生成任务看作从输入到输出的端到端转换过程，因此业界通常采用深度学习模型完成端到端的文本生成。上述方式的优点是文本生成性能提升明显，且相对更容易开发^④，缺点是高度依赖大规模标注数据，模型的可解释性不好。

图2总结了Data2Text任务的两种框架：流水线框架和端到端框架。图2以足球比赛赛事简讯生成任务为例，通过赛事数据页面（见左上角）获得输入数据，然后采用任一框架进行赛事简讯生成，输出结果见右上角文本。

④ 现有多多个深度学习平台（如TensorFlow、Pytorch等）支持各类深度学习模型的复用与开发。

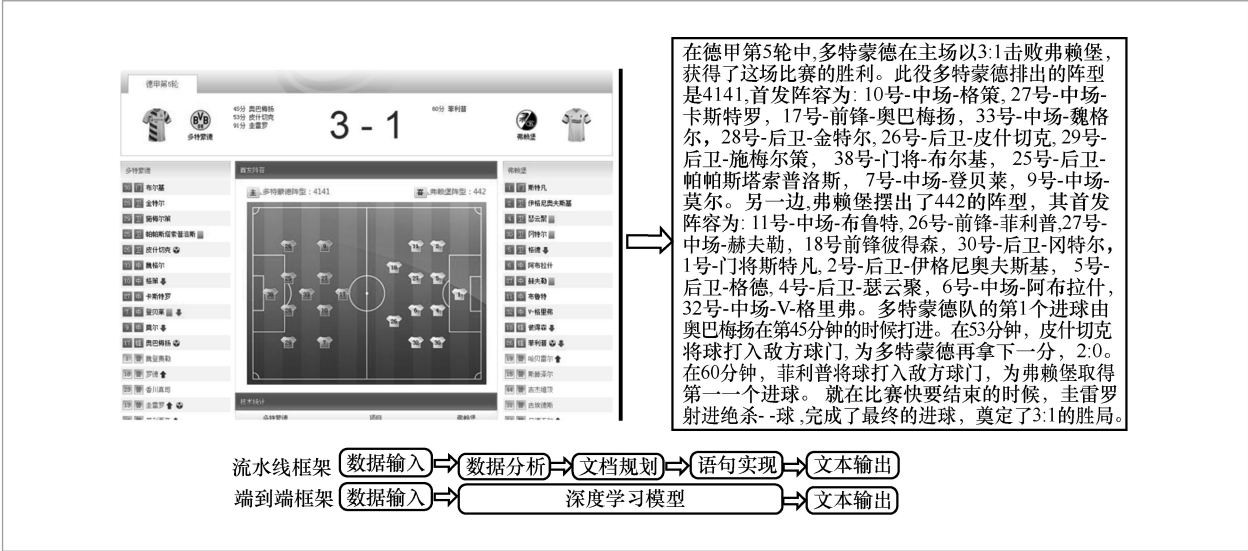


图 2 基于足球比赛数据的赛事简讯生成任务的两种框架（左上角为足球比赛数据展示页面，右上角为输出的简讯文本）

3.2 基于编码器-解码器框架的文本生成方法

尽管不同的深度学习模型被尝试应用于文本生成,但近些年最成功的模型是基于编码器-解码器 (encoder-decoder) 框架的方法。编码器用于对输入数据进行理解和编码,计算数据的语义向量表示,而解码器则以编码器的输出为输入,进行词语序列的解码输出。面向不同的文本生成任务,编码器可采用不同的深度学习模型对不同类型的输入进行编码,包括循环神经网络 (recurrent neural network, RNN)、长短时记忆网络 (long short-term memory, LSTM)、卷积神经网络 (convolutional neural network, CNN)、Transformer网络等。不同文本生成任务的输出均为文本,也即词语序列,因此解码器可采用RNN、LSTM或Transformer网络,这些模型基于已有的词语序列预测下一个词语。具体来说,其先解码输出第一个词语,接着解码第二个、第三个词语,以此类

推,直到输出结束符号为止。

许多文本生成任务可看作序列转换问题,即将输入符号序列转换为输出符号序列。输入符号序列不限于词语,也可包括数值、特殊符号等,例如Data2Text任务输入的符号序列可包括词语、数值、标签等,AMR2Text任务输入的符号序列可包括概念、概念关系标签、特殊符号等。用于求解序列转换问题的模型统称为Seq2Seq模型^[2],其本质上也基于编码器-解码器框架。图3展示了以循环神经网络为基础的Seq2Seq模型框架。图3将输入符号序列(A, B, C, D)转换为输出符号序列(X, Y),其中<BOS>为起始符号,<EOS>为结束符号。

为获得更好的文本生成结果,Seq2Seq模型一般采用注意力机制,使解码阶段能够关注到输入端的序列信息。例如,对于图3的例子,解码第二个符号Y时可能更关注输入符号序列中的符号C,而其他符号A、B、D对Y的生成影响较小。模型会根据当前解码状态与编码器每个状态的相关度计算注意力权重,然后基于注意力权重计算获得关于输入信息的上下文向

量,并用该向量帮助解码。

Transformer模型^[3]同样基于编码器-解码器框架,也适用于处理序列转换问题,可看作一种充分利用注意力机制的特殊Seq2Seq模型。Transformer模型的编码器主要由多头注意力(multi-head attention)、残差连接与层归一化(add & norm)、前向反馈(feed forward)网络等模块构成,为每个输入符号获得一个上下文感知的语义向量表示。解码器与编码器类似,主要由隐码多头注意力(masked multi-head attention)、多头注意力、残差连接与层归一化、前向反馈网络等模块构成。解码器通过多头注意力对编码器输出信息进行有效的访问和利用,同时通过隐码多头注意力访问和利用已经生成的序列信息。Transformer模型在各类文本生成任务上均具有卓越的性能表现,已成为当前文本生成乃至整个自然语言处理领域的代表性和突破性技术之一,具有重要的落地应用价值。

不同于Seq2Seq模型的其他代表性文本生成模型有变分自动编码器(variational autoencoder, VAE)^[4]、生成对抗网络(generative adversarial network, GAN)^[5]、扩散模型(diffusion model)^[6]等,这些模型要么难以训练,要么当前的性能表现仍不理想,难以在实际场景得以应用,需要进一步探索和完善。

3.3 基于预训练语言模型的文本生成方法

预训练语言模型(pretrained language model, PLM)是对海量语料(一般是未标注语料,也可包括标注语料)进行训练得到的语言模型。语料一般不带任何标注信息,因此模型通常采用一

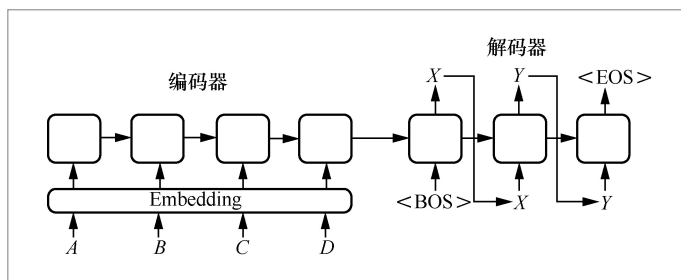


图3 以循环神经网络为基础的Seq2Seq模型框架

个或多个自监督任务作为训练目标,例如基于文本中前面的词语序列预测下一个词语等。预训练语言模型通常由多层基础神经网络模块(如Transformer网络)堆叠而成,参数规模大(从数亿增长到数千亿),需要大量的数据进行训练。不同的预训练语言模型具有不同的架构,通过模型预训练能获得性能卓越的文本编码器(如BERT^[7]、RoBERTa^[8]模型)或文本解码器(如GPT系列模型^[9-12]),或者二者兼得(如BART^[13]、T5^[14]模型)。其中,OpenAI推出的GPT系列模型是最具代表性同时也是到目前为止最成功的文本生成大模型,特别是2022年推出的ChatGPT模型具有令人印象深刻的文本生成能力。其能够根据用户的提示和引导完成各类文本生成任务,例如文本摘要、复述、故事生成、文案生成等,还具有回答问题、生成代码等多种能力,被部分学者视为走向通用人工智能(artificial general intelligence, AGI)的一条可行路径。GPT系列语言模型总结见表1。可以看到,GPT系列模型的发展趋势是首先增加模型参数和训练数据规模使其具备足够强的生成能力,然后考虑利用人类反馈信息优化模型生成更加安全并符合人类价值观的内容。

预训练语言模型的超强能力可以总结为“一种泛化的记忆能力”,通过大规模模型参数以某种隐式方式获取并存储数据中的语言学及事实、常识等各类知识,

表1 GPT 系列语言模型总结

模型名称	发布时间	最大参数规模/条	训练数据	模型架构
GPT	2018年	约1.2亿 (117 M)	BooksCorpus	12层TransformerDecoder
GPT-2	2019年	约15亿 (1 542 M)	WebText (约40 G文本)	48层TransformerDecoder
GPT-3	2020年	约1 750亿	Common Crawl (filtered)、WebText2、Books1、Books2、Wikipedia (共约5 000亿个词)	96层TransformerDecoder
InstructGPT	2022年	约1 750亿	数万提示文本及生成结果标注	GPT-3 + 基于人类反馈的强化学习
ChatGPT	2022年	与GPT-3相当	额外的标注数据 (具体信息未知)	与InstructGPT类似

同时具备一定程度的涌现能力 (emergent ability), 解决相对复杂、困难的任务。目前对于语言大模型内在机理的研究还处于起步阶段, 语言大模型对人们而言仍然像是一个黑盒子, 亟待业界共同努力破解其内在奥秘。

预训练语言模型相当于“巨人的肩膀”, 已成为解决各类文本生成任务的关键支撑, 众多研究试验对比发现, 采用预训练语言模型能大幅提升文本生成效果。业界通常采用如下两种典型方式^⑤利用预训练语言模型帮助特定文本生成任务: 一是面向特定任务的模型微调 (finetuning); 二是对模型进行提示和示例 (prompting & demonstration)。第一种方式基于特定任务的标注数据对预训练语言模型进行进一步训练, 调整优化预训练语言模型的全体参数, 从而取得在特定任务上的更佳效果。例如, 可以基于文本简化数据进一步训练BART模型, 使得BART模型能够更好地应对文本简化任务。与第一种方式不同, 第二种方式不用进一步训练预训练语言模型, 甚至可以不用特定任务的标注数据, 或者只利用极少量的数据样例。该方式高度依赖人工或自动构造的自然语言提示^⑥, 在该提示的引导下, 预训练语言模型能够理解任务并生成相应的结果。例如, 对于文本简化任务, 可以构造这样的提示“请将这句话改写为更简单易懂的句

子: 【*】, 改写结果如下: ”, 然后用待改写的句子替换“【*】”作为GPT-3模型的输入, 模型会生成/续写相应的简化结果。为了进一步提升模型对特定任务的理解, 可以提供一个或多个样例 (包括样例的输入和标注结果), 与提示文本拼接在一起作为模型输入, 使得模型能够通过样例更好地理解特定任务的目标和要求, 从而更好地完成任务。可以看到, 对模型进行提示的方式不需要重新训练模型, 因此能够节约资源且更加方便, 但是要求模型已经具备多任务、少样本甚至零样本文本生成能力。从实际效果来看, GPT-3和ChatGPT能够比较好地支持以该方式完成各类主流文本生成任务, 而对于独特、小众、领域相关的文本生成任务, 可能仍需要采用模型微调的方式才能取得较好的生成效果。

4 智能文本生成面临的挑战

基于深度学习和预训练语言模型的智能文本生成已经展现出强大的能力, 但是仍面临可控性不好、缺乏可解释性、资源消耗大等问题和挑战。首先, 模型无法精准控制生成结果, 导致生成文本的质量时好时坏, 进而影响其在关键场景的落地应用, 这方面的具体问题稍后会展开

⑤ 业界还提出基于任务标注数据对预训练模型的一部分参数或者额外增加的参数进行训练的方法, 但此类方法既不简洁, 效果也不突出, 因此不再赘述。

⑥ 不同的提示可能会导致不同的生成效果, 如何构造好的提示也是值得探讨的问题。

详述。其次,基于深度学习的文本生成模型缺乏可解释性,其工作机理更像是一个黑盒子,不利于错误诊断和分析。关于预训练语言模型的内在机理,业界虽然有各种推测和假设,但目前并没有令人信服的答案。最后,目前的预训练语言模型都非常庞大,训练和使用会耗费巨大的资源(GPU卡、存储空间、能源消耗、资金投入等),与人类社会的绿色环保发展理念相背离,且容易形成垄断,不利于行业生态的健康发展。目前针对GPT系列大模型的压缩和小型化工作极少,且难度很大,轻量级文本生成模型的研制需要得到业界更多的关注和重视。文本生成模型的结果不可控问题具体体现在以下几个方面。

- 内容覆盖性不佳(“写得不够”)。典型的有条件文本生成任务(如文本摘要、Data2Text、AMR2Text等)均要求输出结果能够覆盖输入数据中的重要信息。然而,当前的文本生成模型可能会遗漏输入数据中的某些重要事件/关系/实体等信息,导致出现内容覆盖性不佳的问题。例如,对输入的句子“小丽的爸爸和妈妈来美丽的校园看望她了”进行摘要和压缩,模型可能会生成“小丽的爸爸来校园看望她了”,这个结果文本遗漏了输入句子中的重要信息“小丽的妈妈”,因此内容覆盖性较差。当前解决该方法的方法包括显式利用句法语义知识增强模型编码端对输入数据/文档的理解,也可以考虑首先提取重要信息(如关键词、实体、关系等),然后在模型解码端增加约束,要求模型输出覆盖重要的词语/实体/关系。

- 文本多样性欠缺(“写得雷同”)。目前的文本生成模型倾向于在类似的条件下生成相似的文本表达,使得输出文本欠缺多样性,给读者千篇一律的感受。例如,对不同场次的足球比赛,模型可能会生成

类似的语句“A队X-Y战胜了B队”介绍比赛结果(其中A、B为比赛队伍名,X-Y为比分),这样的赛事报道文本表达雷同,给读者不好的阅读感受。该问题的根本原因在于文本解码算法在每一步均倾向于输出(给定上下文条件下)神经网络模型计算得到的概率最大的词语。业界解决该问题的方法主要是在解码过程中简单引入随机性(如top- k 、top- p),或者通过采样不同的随机向量帮助多样性文本生成。这些方法虽然能够部分改善这个问题,然而会引起文本基本质量(如流畅性)下降。

- 信息保真性不好(“写得有错”)。当前文本生成模型的输出结果中通常会出现关键信息与输入数据或客观世界不一致或相互冲突的情况,也会出现某些无法验证的信息。例如ChatGPT在用户的引导下会输出“曹雪芹的丈夫名叫高鹗”这样与客观事实不符的文本。文本摘要模型会在生成结果中用错实体、数量、性别等关键信息(例如原文中为“张三”,而在摘要结果中为“李四”)。这个问题严重影响了文本生成模型的应用。可以看到,由于模型在训练阶段见识过大量数据,并记忆各类信息,这些信息仍有可能在测试使用阶段出现在模型输出中,也有可能当前输入信息进行某种形式的组合作为输出。因此,业界主要考虑对模型输出增加语义约束从而提升文本生成结果的语义保真性,或者利用强化学习手段进行模型优化。此外,业界也探索利用纠错模型来修正文本生成结果的错误,提升信息保真性。

- 内容安全性不足(“写得有害”)。目前的文本生成模型在一些特定的条件或提示下会生成不安全、不友好的文本内容,例如侮辱性言论、种族歧视言论、性别歧视言论等不当言论,也可能会生成诸如“请食用从山上采到的彩色蘑菇”这类威胁人身安全的语言。业界设计了多个

可控文本解码算法（例如DEXPERTS^[15]）在解码过程中控制输出文本的属性类别。ChatGPT则通过基于人类反馈的强化学习手段，利用人工标注数据优化模型，尽量让模型与人类价值观对齐，在实际测试中具有良好的表现，但受限于标注数据的规模和领域覆盖性，以及强化学习方法自身的能力，内容安全性不足的问题并没有得到真正的解决。

- 文本生动性不够（“写得不美”）。智能文本生成的最终目标之一是能够自动生成更加生动和形象的文本，让读者感受到语言之美，体会到文字感染力。实现这个目标的主要方式包括巧用修辞手法（如双关、隐喻、比喻、夸张等）、引经据典、巧用幽默和词语等。然而，当前的文本生成模型不具有上述写作能力，仅能生成相对平实、直白的语言。业界已有少数工作分别聚焦在不同类型修辞语言的生成、引文推荐、简单幽默的生成，但任务定义和实现方法均比较初级，距离实现文本生动性的目标还有很大的差距。

- 篇章连贯性不强（“写得不通”）。当前的文本生成模型通常具有能力较强的解码器，能够生成流畅的语句，但是在生成长文本时会出现句间连贯性的问题，语句之间可能存在连贯性较差或相互矛盾的情况，也会存在篇章逻辑方面的问题。例如，文本生成大模型会生成“……我不是你的主人，我是你的主人……”这样前后矛盾的文本，这样的结果会严重影响读者的阅读和理解。业界已尝试利用强化学习手段提升连贯性，或利用内容规划信息指导文本生成，但是长文本的连贯性问题仍有待深入探索。

5 总结与展望

本文对智能文本生成的应用现状、主要方式、主流方法及面临的挑战进行了全

面的阐述。智能文本生成技术不仅能用于智能写作，还能用于解决其他几乎所有自然语言处理任务，即“万物皆可生成”。从近几年的研究趋势可以看到，绝大部分自然语言处理任务（包括智能问答、信息抽取、句法分析、语义分析等）可以通过序列转换模型或预训练语言模型加以解决，并能取得突出的试验效果。因此，更好的文本生成方法/模型将能够促进整个自然语言处理领域的发展。

展望未来，除了继续探索可控性文本生成方法之外，业界还可以探索人在回路的文本生成，使模型可以充分、持续利用人类与模型交互过程中所提供的显式或隐式反馈信息（例如基于模型生成结果完成下游任务的成功率和时间开销等）实现自我进化。此外，尽管自然语言处理任务通过文本生成方式基本实现了大一统的解决框架，但是能否跨越不同模态，研制同时支持图像、视频、语音和文本的大一统模型，也是一个极具挑战性的问题，业界针对这个问题已有一些尝试，但目前效果还不能令人满意。最后，文本质量评估是衡量和驱动文本生成技术研究的“尺子”和“利器”，然而当前文本质量自动评估方法可靠性不高，人工评价方法可重复性低，业界亟须探索更可信的自动评价指标与可重复的人工评价框架。

参考文献:

[1] REITER E. An architecture for data-to-text systems[C]//Proceedings of the 11th European Workshop on Natural Language Generation. New York: ACM Press, 2007: 97-104.

[2] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]//Proceedings of the 27th International Conference on Neural

- Information Processing Systems. New York: ACM Press, 2014: 3104–3112.
- [3] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 6000–6010.
- [4] KINGMA D P, WELLING M. An introduction to variational autoencoders[J]. Foundations and Trends® in Machine Learning, 2019, 12(4): 307–392.
- [5] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139–144.
- [6] LI X L, THICKSTUN J, GULRAJANI I, et al. Diffusion-LM improves controllable text generation[J]. arXiv preprint, 2022, arXiv: 2205.14217.
- [7] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint, 2018, arXiv: 1810.04805.
- [8] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[J]. arXiv preprint, 2019, arXiv: 1907.11692.
- [9] RADFORD A, NARASIMHAN K, Salimans T, et al. Improving language understanding by generative pre-training[Z]. 2018.
- [10] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019, 1(8).
- [11] 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 Press, 2020: 1877–1901.
- [12] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[J]. arXiv preprint, 2022, arXiv: 2203.02155.
- [13] 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: 7871–7880.
- [14] COLIN R, NOAM S, ADAM R, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research, 2020, 21: 5485–5551.
- [15] LIU A, SAP M, LU X M, et al. DExperts: decoding-time controlled text generation with experts and anti-experts[J]. arXiv preprint, 2021, arXiv: 2105.03023.

作者简介



万小军 (1979–), 男, 博士, 北京大学王选计算机研究所博士生导师, 主要研究方向为自动文摘与文本生成、情感分析与语义计算、多语言与多模态NLP等。曾担任计算语言学重要国际期刊*Computational Linguistics*编委、国际会议EMNLP 2019程序委员会主席, 现任CCF-NLP专委会秘书长、中国中文信息学会理事与NLGIW专委会副主任、TACL与ARR执行编委、NLE编委、JCST编委, 多次担任相关领域重要国际会议(ACL、NAACL、EMNLP、EACL、AAACL)高级领域主席或领域主席。荣获ACL 2017杰出论文奖、IJCAI 2018杰出论文奖。研制推出多款AI写作机器人, 如小明、小南、小柯等, 应用于多家媒体单位。

收稿日期: 2023-02-13

通信作者: 万小军, wanxiaojun@pku.edu.cn

基金项目: 国家自然科学基金资助项目(No.62161160339); 国家重点研发计划资助项目(No.2021YFF0901502)

Foundation Items: The National Natural Science Foundation of China(No.62161160339), The National Key Research and Development Program of China(No.2021YFF0901502)