技术进展

机器写作：让计算机掌握文字创作的本领

万小军
北京大学

关键词：机器写作　自然语言生成

机器写作（又称自然语言生成[1]）是自然语言处理领域的重要研究方向之一，也是近期研究热点。我们希望计算机同时具有读与写的能力，除了掌握阅读和理解语言文字的本领之外，还能够掌握文字创作的本领，从而像人类一样写出高质量的各类文字作品，例如新闻资讯、报告、诗歌、小说、作文等。

机器写作在传媒、出版、文娱、广告等多个行业均具有广阔的应用场景。欧美等地较早就创建了多家专注于机器写作、技术应用的公司，例如ARRIA[1]、AI[2]、NarrativeScience[3]等，这些公司基于行业数据生成行业报告或新闻报道，从而节省了大量的人力。同时，不少国外知名媒体单位纷纷采用机器写作技术进行新闻稿件创作，替代编辑与记者的部分工作。例如，2006年美国汤姆森公司开始用机器人撰写金融新闻，2014年美联社全面利用机器人WordSmith（AI公司的写作引擎）进行写作。近几年，机器写作在国内也逐渐受到业界的重视，包括今日头条、腾讯、百度、360等各大互联网公司以及新华社、南方都市报、第一财经等传统媒体单位均开展了机器写作技术的研究与应用。他们推出了Xiaomingbot、DreamWriter、快笔小新、小南、DT稿王等多款写作（写稿）机器人。其中，Xiaomingbot是我们与今日头条在2016年7月联合推出的写稿机器人。它能够针对各类体育赛事撰写稿件，包括短篇消息与长篇通讯，目前累计撰写的新闻稿件已达万篇，获得数千万的阅读量。图1展示了Xiaomingbot的页面与新闻实例。

通过上述应用表明，机器写作不再属于纸上谈兵，而是已经对人类工作和生活产生了重大影响。与人类作者相比，机器写作具有效率高、实时性好、覆盖性强、无偏见等优点，同时，据今日头条的线上测试表明，机器人撰写的新闻稿件的阅读率与人工稿件的阅读率基本相同，这说明机器稿件的质量是不错的，能够被广大用户所接受。

几种不同的机器写作方式

计算机不能凭空写作，而必须根据所输入的数据与素材进行创作。根据不同类型的输入，计算机一般采用不同的写作方式进行创作。

[1] https://www.arria.com
[2] http://automatedinsights.com
[3] http://www.narrativescience.com

(a) (b)

图1 (a)Xiaomingbot的界面 (b)Xiaomingbot自动撰写的新闻实例

原创：计算机根据输入的结构化数据（报表、RDF 数据等）进行文字创作。该方式能够生成原创稿件，是目前机器写作的主要方式，适用于天气预报、医疗报告、赛事简讯、财经报道等文本的生成。

数据到文本生成的框架如图 2 所示。该框架由英国阿伯丁大学的雷特 (Ehud Reiter) 在三阶段流水线模型的基础上提出[2]。其中信号分析模块 (signal analysis) 通过利用各种数据分析方法检测输入数据的基本模式，输出离散数据模式；数据阐释模块 (data interpretation) 通过对基本模式和输入事件进行分析，推断出更加复杂和抽象的消息，同时推断出它们之间的关系，之后输出高层消息以及消息之间的关系；文档规划模块 (document planning) 分析决定哪些消息和关系需要在文本中提及，同时确定文本的结构，最后输出需要提及的消息以及文档结构；微规划与实现模块 (microplanning and realisation) 根据选中的消息及结构通过自然语言生成技术输出

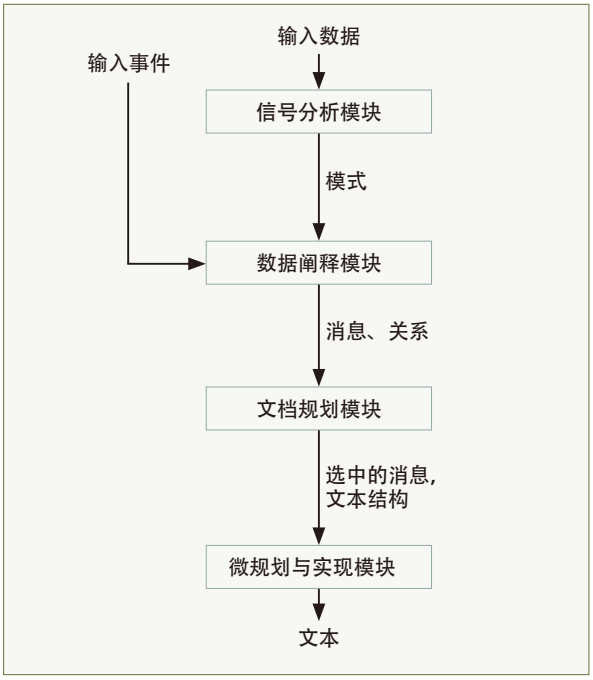

图2 数据到文本的生成系统的一般框架[2]

最终的文本，主要涉及对句子进行规划以及句子实现，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。

在实际应用中，可采用模板制作和填充的方式实现数据到文本的生成。模板制作依赖于领域专家的写作经验，或者从大量平行语料中进行模板的自动学习与归纳。一旦模板制作完成，稿件的写作过程就很简单，只需要将相关数据填充到模板中即可。这种方式所生成的稿件准确性高、可读性强；然而，由于模板比较固定，所生成稿件的模式基本相同，多样性较差。当然，如果模板库足够丰富，稿件多样性少的问题也可以得到解决，但制作模板比较耗时耗力。另一个严重的问题在于，不同领域的机器写作依赖于不同的数据和不同的模板，而模板的领域迁移性很差，这也导致目前机器写作应用难以实现跨领域迁移，面向新领域需要二次开发。

基于深度学习技术进行数据到文本的生成则不依赖于模板或规则，而是直接基于平行语料学习得到端到端的生成模型。然而，这样的写作方式虽然在研究上取得了一定的进展，但目前并不能保证所生成稿件的准确性与可读性，难以达到对稿件的高质量要求。此外，基于深度学习的生成模型需要大量（一般数万以上）的平行语料，而目前在很多领域又较难获取。

二次创作：计算机根据已有的文字素材（例如，已经发表的新闻）进行二次文字创作。该方式能够基于已有稿件创作出不一样的稿件，例如为一篇新闻生成摘要，对多篇相关新闻进行综述，对一篇新闻进行文字改写等。

二次创作主要依赖于两类自然语言处理技术：自动文摘与文本复述。自动文摘用于对单篇文本或多篇文本进行内容提炼与综合，形成摘要或综述。Xiaomingbot 写稿机器人就利用了基于机器学习的自动文摘技术对平均长达 5000 字的赛事直播文字进行筛选与融合，形成长达千字的赛事报道[3]。需要指出的是，多文档自动文摘比单文档自动文摘更具有挑战性，原因在于不同文档内容的冗余性、片面性与弱连贯性。因此，对多篇新闻报道进行长篇综述的生成极其困难，我们在这方面进行了一些尝试，提出基于段落排序与融合的方法，取得了一定的效果[4]。文本复述则用于对现有文字进行改写，在主题与意思基本不变的前提下产生另一种文字表述，从而避免原文照抄，也可实现文本风格化的目的。例如，可以将规范的书面用语改写为活泼的网络用语，也可将奥巴马的语言风格转换成特朗普的风格。可以将文本复述看作是一种单语言机器翻译问题，因此在平行语料充足的前提下，各种统计机器翻译方法（包括神经网络机器翻译）均可应用于此。但现实中却难以获得大规模的此类平行语料，因此针对文本复述的研究需要另辟蹊径，最新的研究主要集中在如何有效地利用少量的平行语料和大规模的非平行语料进行复述模型的学习。

混合创作：计算机可以结合原创与二次创作两种方式进行文字创作，稿件中的一部分内容从结构化数据中直接生成，而另一部分内容则从已有文本中进行提炼或改写得到。混合创作能够生成内容更加丰富、形式更加多样的文本。

深度学习在机器写作中的应用

最近几年，无论是从数据到文本的生成，还是自动文摘与文本复述，都不可避免受到深度学习热潮的波及。编码器 - 解码器框架 (encoder-decoder) 则是应对各类机器写作任务的通用框架，该框架采用一个编码器对输入的数据或文本素材进行语义编码，获得隐式语义向量，随之采用一个解码器逐词进行解码，直到遇到结束符。针对不同的输入，可采用不同的编码方法：针对结构化数据输入，可对每条数据记录进行编码然后进行拼接；针对文本输入，则可利用 RNN/LSTM/GRU、CNN 等模型进行编码。解码器则主要基于 RNN/LSTM/GRU 模型，采用贪心搜索或柱搜索来获得词语序列，即结果文本。图 3 展示了以文本作为输入的编码器 - 解码器框架图，即序列到序列生成模型 (Seq2Seq)。该框架可进一步集成注意力机制 (attention mechanism)、拷贝机制 (copy mechanism) 等来改善生成效果。

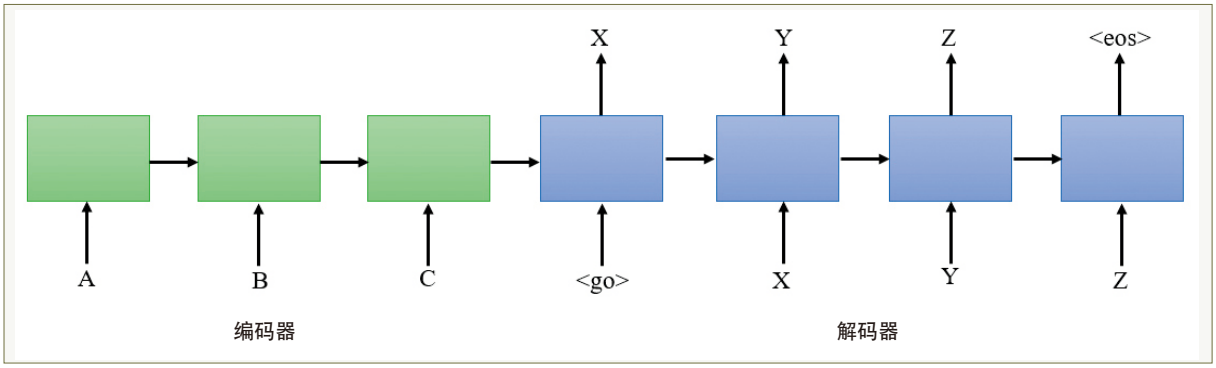

图3 编码器-解码器框架（以文本输入为例）

上述编码器 - 解码器框架的好处在于能以一种端对端的方式进行文本生成，避免了传统框架中复杂的多阶段生成流程，让文本生成过程变得简单。该框架的另一个好处在于可以提高生成文本的多样性，让同类输入甚至同一输入获得具有不同语言表达的输出文本。

然而，上述框架也存在不足之处：一是需要大规模的平行语料进行训练，而很多应用场景却缺乏此类平行语料，因此有必要探索小数据下的学习机制来解决这个问题；二是生成文本中信息与数据的准确性、文本的可读性不能完全得到保证（尤其是长文本），而不少应用场景不能容忍文本质量上的瑕疵，因此需要考虑结合更先进的技术来提高所生成文本的质量，业界目前已经逐步尝试利用强化学习、生成式对抗网络 (GAN) 等手段来帮助文本生成，这类技术手段能够针对整个文本的某个质量目标（例如 BLEU 指标）直接进行优化。

机器写作展望

机器写作无论是在研究上还是在应用上都取得了明显的进展，但也面临不少难点。除了上文提到的平行语料缺乏、领域迁移性差等问题之外，还存在难以客观评价的问题。目前机器写作的客观评价指标一般为 BLEU 和 ROUGE，这两个指标用来计算生成文本与参照文本之间的词语重叠程度。然而，文章的写作方式可以有千万种，每个作者都可以根据同一命题写出内容不一样但质量都很高的文章。那么，目前为生成文本只提供一两个参照文本的评价方式显然不合理，但现实却又只能如此。当然，我们可以采用人工进行主观评价，然而人工评价耗时耗力，同时会受到外在因素的干扰。因此，未来有必要设计更合理的客观评价指标，这对机器写作领域的发展会起到不可估量的推动作用。

机器写作除了用于撰写新闻、报告等应用型文本之外，近几年还被用于创作古诗、现代诗、散文等文学作品，例如微软小冰、清华九歌等系统，能够分别创作现代诗和古诗，在文字形式上的总体效果还是不错的，但在意境上有所欠缺。我们也尝试了基于自动文摘的方式进行散文的自动“拼凑”，经中学教师评分能够取得不错的分数 [5]。

在未来几年，机器写作将在更多行业和领域得到应用，从而节省大量的人力，体现更大的价值。同时，随着数据的逐步累积和模型的逐步完善，基于深度学习的机器写作将会取得显著的进展。 ■

万小军
CCF 专业会员。北京大学计算机科学技术研究所研究员。主要研究方向为文本挖掘与自然语言处理、自动文摘与文本生成、情感分析与语义计算等。
wanxiaojun@pku.edu.cn

参考文献

[1] Gatt A, Krahmer E. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation[OL]. (2017-12-19). arXiv:1703.09902v3.

[2] Reiter E. An architecture for data-to-text systems[C]// Proceedings of the Eleventh European Workshop on Natural Language Generation. ACL Press, 2007: 97-104.

[3] Zhang J, Yao J G, Wan X. Towards Constructing Sports News from Live Text Commentary[C]// Proceedings of the Meeting of the Association for Computational Linguistics. 2016:1361-1371.

[4] Zhang J, Wan X. Towards Automatic Construction of News Overview Articles by News Synthesis[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language. ACL Press, 2017: 2111-2116.

[5] Li L, Wan X, Yao J and et al. Leveraging Diverse Lexical Chains to Construct Essays for Chinese College Entrance Examination[C]//Proceedings of the 8th International Joint Conference on Natural Language Processing. 2017.