OmniThink: Expanding Knowledge Boundaries in Machine Writing through Thinking

2025年6月1日　2 分钟阅读

#machine_writing

> *本文介绍了OmniThink: Expanding Knowledge Boundaries in Machine Writing through Thinking。*

我为什么要阅读这篇文章?

从STORM到Co-STORM的演进，以及随后出现的如OmniThink等批判和提议，突显了人工智能写作领域一个关键且持续的挑战：超越单纯的信息聚合和貌似合理的文本生成，迈向真正的知识综合、深度理解和新颖见解的产出。

1、STORM专注于**结构化的预写过程**，Co-STORM则加入了**人类协作和动态知识图谱**，以改善覆盖面和偶然性发现，OmniThink随后批判了像STORM这样的系统在深度、新颖性和知识组织方面的局限性，并提出了带有**信息树和概念池**的"慢思考"方法。

2、STORM自动化了研究和提纲挈领的过程。Co-STORM将人类引入了这个自动化循环。OmniThink则试图在人工智能自身内部构建更复杂的内部认知模型（信息树、概念池）。每一步都旨在解决前一步的不足。

3、STORM解决了基础RAG的肤浅性问题。Co-STORM通过利用人类互动来解决潜在的信息鸿沟和"未知未知"问题。OmniThink则针对即使在像STORM这样的高级RAG系统中仍然存在的缺乏真正深度、新颖性和有效知识结构化的问题。

4、要生成真正具有深度和广度的"专业"文章，不仅需要查找和排列事实，还需要将它们综合成新的理解，识别新颖的联系，并以深度连贯的方式构建知识。从STORM到OmniThink的历程，展示了人工智能研究界为赋予机器这些更高级认知能力所做的努力，从信息处理迈向知识创造。已识别的局限性（偏

目录

文章信息

字数

阅读时间

发布时间

更新时间

标签

#machine_writing

见、冗余、缺乏新颖性）正是阻碍当前人工智能持续达到人类专家写作最高标准的障碍。

更多讨论在： https://gemini.google.com/app/d98d771d0864e8e3

5、这一系列进展显示出在尝试复制或支持复杂人类写作认知过程方面日益增长的复杂性。

6、OmniThink 的提出，标志着从信息处理到知识创造的转变。

7、同时，在RAG领域的研究中，也出现了很多新的方法，这些方法在一定程度上解决了RAG的局限性，可以结合起来使用。

OmniThink Expanding Knowledge Boundaries in Machine Writing through Thinking

本文提出了一种新的机器写作框架——OmniThink，旨在通过模拟人类的思考过程来扩展知识边界，从而提高生成文章的质量。该框架分为三个阶段：信息获取、概念引导的提纲构建和文章合成。在信息获取阶段，通过建立信息树和概念池来逐步收集和组织相关信息，这一过程中结合了系统的搜索与反射操作以保证信息的深度和广度。提纲构建阶段利用已经提炼的概念池指导文章的结构布局，确保内容的逻辑性和连贯性。最后，在文章合成阶段，利用事先准备的信息树进行内容抽取和拓展，同时利用语言模型完成最终文本的生成。实验结果表明，与现有方法相比，OmniThink能够显著提升生成文章的知识密度，并且保持良好的语义连贯性和深度，为长期的文本生成任务提供了一个新方向。此外，文中还提出了一个新的评价指标——**知识密度（KD）**，用以衡量文章中有效信息的比例，弥补了现有评价体系的不足。

关键要点

研究提出了一种新的机器写作框架——OmniThink，通过模拟人类的思考过程来提高文章的知识密度和质量。

OmniThink的核心思想是模仿学习者逐渐加深对复杂主题的理解以扩展知识边界的过程。

OmniThink引入了两个创新组件——信息树和概念池，用于模拟收集信息和结构化认知的过程。

实验结果表明，OmniThink提高了生成文章的知识密度，同时不影响关键指标如连贯性和深度。

OmniThink为未来的长篇文章生成研究提供了新方向，并提出了一个新的度量标准——知识密度（KD）。

论文主要内容

以下是对这篇论文的十个问题及其解答，内容来自总结论文的摘要和主要内容。

1. 这篇论文的主题是什么?

这篇论文提出了一个名为 OmniThink 的机器学习写作框架。该框架旨在通过模仿人类的迭代式扩展和反思过程来提升机器写作的知识边界。具体而言，它专注于开放领域长文本生成任务。

2. 为什么这个主题很重要?

机器写作（特别是使用大型语言模型进行长文本生成）通常依赖于检索增强生成（RAG）。然而，现有的 RAG 方法往往局限于模型的预定义范围，导致生成的内容信息不丰富。具体来说，简单的检索信息可能缺乏深度、新颖性，并且存在冗余。这会负面影响生成文章的质量，使其变得浅显、缺乏原创性和重复。人类写作可以自然地避免这些问题，这可以通过认知科学中的反思实践理论来解释。人类作者通过持续反思先前收集的信息和个人经验，重组、过滤和提炼认知框架，从而迭代调整写作方向和思维路径，最终生成更深刻、细致和原创的内容。因此，提升机器写作的质量、克服冗余和缺乏新颖性等问题，使其更接近人类的思考和写作过程，具有重要意义。

3. 论文的主要贡献或提议是什么?

论文的主要贡献是提出了 OmniThink 框架。OmniThink 是一个模仿人类慢思考过程的新型机器写作框架。它的核心思想是模拟学习者逐步加深对复杂主题理解的认知行为，以扩展知识边界。论文还提出了一个新的评估指标：**知识密度 (Knowledge Density, KD)**，用于衡量文章中有用信息的比例。此外，论文还从新的知识边界视角分析了当前长文本生成方法的挑战，探究了 OmniThink 有效性的潜在因素，并为未来长文本生成研究提出了新的方向。

4. 提议的方法是如何工作的?

OmniThink 框架被分为三个主要步骤：

> **信息获取 (Information Acquisition)**：这个阶段主要通过持续的**扩展 (Expansion)** 和**反思 (Reflection)** 来形成一个信息树 (Information Tree) 和概念池 (Conceptual Pool)。这些构成了后续大纲构建和文章生成的基础。这个过程是迭代进行的。
>
> **概念引导的大纲构建 (Concept-guided Outline Structuring)**：在收集并组织了多样化的信息后，OmniThink 进入大纲构建阶段。它首先创建一个大纲草稿，然后利用概念池中的内容对草稿进行提炼和关联，形成最终的大纲。概念池代表了 LLM 对主题扩展后的认知边界。

文章撰写 (Article Composition)：完成大纲后，OmniThink 并行地为大纲的每个部分撰写内容。在撰写某个部分时，它会根据该部分的标题和子标题从信息树中检索最相关的文档，并基于检索到的信息生成带引用的内容。所有部分生成后会被拼接起来形成文章草稿，然后模型会进一步处理拼接后的文章，删除冗余信息，形成最终的文章。

5. 该方法的关键组成部分是什么?

OmniThink 引入了两个创新的组件来模拟人类迭代学习过程中收集信息和构建认知的过程：

信息树 (Information Tree)：用于模拟收集信息的层次结构。它从一个基于输入主题的根节点开始初始化。在信息获取阶段，OmniThink 分析信息树的叶子节点，并根据当前概念池确定需要深入扩展的领域或方向。然后为每个叶子节点生成子节点代表特定方面或子主题，并检索相关信息存储在这些子节点中，从而丰富信息树的层次结构。

概念池 (Conceptual Pool)：用于模拟认知框架。信息树根节点的初始信息会被分析提取，形成初步的概念池，作为 OmniThink 对主题的基础认知，并指导后续的扩展过程。在每次扩展后，OmniThink 会反思信息树叶子节点中新检索到的信息，分析、过滤和综合核心见解，将其融入概念池中，从而更新和丰富认知框架。概念池的更新代表了 LLM 对主题扩展后的认知边界。

信息获取阶段的**扩展**和**反思**是持续进行的，直到收集到足够的信息或达到预设的最大检索深度。这个迭代过程通过信息树和概念池的持续扩展，逐步扩展了信息的边界和认知的边界。

6. 该方法是如何评估的?

论文使用了 WildSeek 数据集进行评估。选取了代表性的基线方法进行比较，包括 RAG、oRAG、STORM 和 Co-STORM。评估使用了多种指标：

文章质量评估：使用 Prometheus2 模型自动评分，衡量 **相关性 (Relevance)**、**广度 (Breadth)**、**深度 (Depth)** 和 **新颖性 (Novelty)**。

信息丰富度：衡量**信息多样性 (Information Diversity)** (网页间余弦相似度差异) 和**知识密度 (Knowledge Density, KD)**。知识密度定义为文章中原子知识单元中唯一有用信息量与文本总长度之比。

大纲质量评估：衡量**内容引导性 (Content Guidance)**、**层次清晰度 (Hierarchical Clarity)** 和**逻辑连贯性 (Logical Coherence)**。大纲评估使用了基于 GPT-4o-08-06 的 Prometheus2 框架。

人工评估：邀请受过良好教育的志愿者对 OmniThink 和 Co-STORM 生成的文章在相关性、广度、深度和新颖性四个方面进行比较评分。

Unique URL 分析：比较不同方法检索到的唯一 URL 数量，以此衡量信息边界的扩展。

处理时间分析：记录每种方法的平均运行时间。

边界分析：可视化信息边界（使用 PCA 将检索信息映射到二维平面）和通过与 oRAG-Plus（增加检索网页数量的 oRAG）的比较来分析认知边界。

扩展与反思分析：通过消融实验（w/o E&R）和替换扩展/反思模型（使用性能较低的模型）来分析扩展和反思对各项指标的影响。

7. 主要的评估结果是什么?

文章生成：在 Prometheus2 的自动评估中，OmniThink 在所有四个评分标准（相关性、广度、深度、新颖性）上都表现出色，尤其在新颖性方面最为突出。在知识密度方面，OmniThink 也比现有基线方法更具优势。在信息多样性方面，OmniThink 也表现最好.

大纲生成：OmniThink 在内容引导性、层次清晰度和逻辑连贯性方面取得了优越的性能。这归因于概念池的设计使其对主题有了更全面和多样的理解。

人工评估：人工评估结果显示，OmniThink 的平均表现优于当前最强的基线 Co-STORM，尤其在广度指标上有显著提升。尽管自动评估显示新颖性有较大提升，但人工评估中的优势不明显，这表明当前自动评估与人类判断可能尚不完全一致。

Unique URL 分析：OmniThink 检索到的唯一 URL 数量远多于其他方法，表明它可以访问更广泛的多样化网络内容，从而生成更具创新性和深度的文章。

边界分析：可视化结果显示 OmniThink 拥有最大的检索范围，确实通过信息树和概念池扩展了信息边界。与 oRAG-Plus 的比较表明，即使有大量信息，没有概念池的引导（认知边界），LLM 也难以有效利用信息，有时甚至表现更差。

扩展与反思分析：消融实验表明，没有扩展与反思 (w/o E&R) 的 OmniThink 在所有指标上都表现更差，尤其是在信息多样性和新颖性方面。进一步分析表明，反思对于新颖性更为重要，因为它赋予模型重新评估和内省现有知识的能力，并以促进更多样和广阔想法出现的方式整合信息，这类似于认知边界的定义。扩展在知识密度、广度和深度方面更为重要，因为它设定了模型后续信息检索的轨迹，使模型更能利用检索到的信息扩展信息边界，从而增强内容的相关性和知识密度。

8. 论文的局限性是什么?

论文指出了几个局限性：

> 当前工作仅限于搜索和文本生成，开放领域中大量的多模态信息（如图像、视频等）尚未使用。
>
> 没有考虑文本生成的个性化语言风格。因此，生成的文本趋向于学术性质，可能不太适合普通用户的阅读偏好。

9. 该方法与之前的工作有何关系?

现有的机器写作方法通常依赖于检索增强生成 (RAG)。早期的 RAG 方法依赖于固定的搜索策略，缺乏生成多样性，导致对主题的探索不彻底，理解碎片化和不完整。STORM 和 Co-STORM 提出了角色扮演的方法来扩展视角，从多个角度收集信息，从而拓宽信息空间。然而，这些方法仍然受限于扮演的角色范围，难以生成深度内容和突破自身的知识边界。 OmniThink 的不同之处在于，它不依赖固定的搜索策略或角色扮演的有限视角。它模拟人类的慢思考过程，通过信息树和概念池模拟人类获取知识和更新认知框架的过程。通过持续的扩展和反思，它综合并更新其认知框架，然后进一步检索额外的信息。这与之前仅通过单次查询或分解为多个子查询来满足信息需求的方法不同。

10. 未来的研究方向是什么?

未来的研究方向包括：

> 探索更先进的机器写作方法，将更深层次的推理、角色扮演和人机交互相结合。
>
> 解决当前的局限性，例如利用开放领域中的多模态信息。
>
> 考虑个性化的语言风格，使生成的文本更适合不同用户的阅读偏好。

附录：OmniThink vs Co-STORM

OmniThink 和 Co-STORM 都是用于开放领域长文本生成的机器写作方法。它们都旨在提高机器生成文章的质量，特别是解决现有方法中信息不丰富、冗余和缺乏新颖性的问题。然而，它们在核心方法和实现上有所不同。

Co-STORM：基于角色扮演和用户参与的方法

> **方法概述：** Co-STORM 是早期改进开放领域长文本生成质量的工作之一。它引入了一种用户参与式的圆桌讨论机制，以增强检索到的信息的多样性。Co-STORM 和 STORM 都提出了角色扮演的方法来拓宽视角，从多个角度收集信息，从而扩展信息空间。

局限性： 尽管 Co-STORM 通过角色扮演拓宽了信息空间，但这种方法仍然受限于所扮演角色的范围，难以生成深度内容并突破其自身的知识边界。检索到的信息可能缺乏深度和新颖性，并存在冗余。

评估设置： 在论文的评估中，Co-STORM 的实现移除了模拟的人类参与部分。在进行大纲质量评估时，由于 Co-STORM 原本没有中间大纲生成步骤，其大纲是从最终文章中提取的。

OmniThink：模仿人类迭代思考的方法

方法概述： OmniThink 是论文中提出的新型机器写作框架，它模仿了人类迭代式扩展和反思的慢思考过程。其核心思想是模拟学习者逐步加深对复杂主题的理解，以此扩展知识边界。OmniThink 的流程分为信息获取、概念引导的大纲构建和文章撰写三个主要步骤。

关键组成部分与工作原理：

信息获取： OmniThink 通过持续的**扩展 (Expansion)** 和**反思 (Reflection)** 来构建**信息树 (Information Tree)** 和**概念池 (Conceptual Pool)**。信息树模拟了信息收集的层次结构，概念池模拟了认知的框架。初始信息形成初步概念池，指导后续扩展。在每次扩展后，OmniThink 反思新信息并将其融入概念池，更新认知框架。这个迭代过程持续进行，扩展信息和认知边界。

概念引导的大纲构建： OmniThink 利用收集到的多样化信息和更新后的概念池来提炼和关联大纲草稿，形成最终大纲。概念池代表了 LLM 扩展后的认知边界。

文章撰写： OmniThink 并行地为大纲各部分撰写内容，根据标题从信息树中检索最相关的信息并带引用生成内容。完成后进行后处理以删除冗余信息。

优势与能力：

OmniThink 不依赖固定的搜索策略或有限的角色扮演视角，而是通过模拟人类思考过程来获取和组织知识。

通过信息树和概念池，OmniThink 扩展了信息边界和认知边界。

消融实验表明，扩展和反思对于提升文章质量至关重要。反思对于提升新颖性尤为重要，因为它促进了多样和广阔思想的出现，类似于认知边界的扩展。扩展则在知识密度、广度和深度方面更为重要，因为它设定了更精确有效的信息检索方向。

OmniThink 检索到的唯一 URL 数量远多于 Co-STORM 等方法，这表明它可以访问更广泛的多样化网络内容，有助于生成更具创新性和深度的文章。

与增加检索网页数量的 oRAG-Plus 相比，OmniThink 的表现更好，这说明即使信息量大，没有概念池的引导（认知边界），

模型也难以有效利用信息。

OmniThink 引入了新的知识密度 (Knowledge Density, KD) 指标，衡量文章中有用信息的比例。

评估结果对比

根据论文的评估结果：

文章生成（自动评估）： 在 Prometheus2 的自动评估中，OmniThink 在相关性、广度、深度和新颖性这四个指标上均表现出色，尤其在新颖性方面最为突出。在知识密度和信息多样性方面，OmniThink 也比包括 Co-STORM 在内的现有基线方法更具优势。Co-STORM 在自动评估中被认为是综合表现最好的基线方法。

大纲生成： OmniThink 在内容引导性、层次清晰度和逻辑连贯性方面取得了优越的性能，这得益于概念池使其对主题有了更全面和多样的理解。由于 Co-STORM 的大纲是从最终文章中提取的，其大纲评分相对较低。

人工评估： 在人工评估中，OmniThink 的平均表现优于 Co-STORM，特别是在广度指标上有显著提升（提升 11%）。尽管自动评估显示新颖性有较大提升，但人工评估中的优势不明显，这可能表明当前的自动评估与人类判断存在差异。

处理时间： OmniThink 的平均处理时间略高于 Co-STORM 和 STORM，但作者认为这些成本是值得的，因为它提高了文本质量。

总结

Co-STORM 通过角色扮演等方式尝试拓宽信息检索的范围，是现有 RAG 方法的改进。而 OmniThink 则通过模仿人类的迭代式“慢思考”过程，引入信息树和概念池来模拟信息的获取和认知的更新。这种机制使得 OmniThink 不仅能访问更广泛和多样化的信息来源（信息边界），更能通过反思和概念池有效地组织和利用这些信息，深化对主题的理解（认知边界）。因此，OmniThink 在自动评估和人工评估中通常表现优于 Co-STORM 等基线方法，尤其在文章的新颖性和知识密度方面具有优势。尽管 OmniThink 的处理时间可能稍长，但其提升的文章质量被认为是值得的。

参考

github: https://github.com/zjunlp/OmniThink

paper: https://arxiv.org/pdf/2501.09751

My NotebookLm Link

分享这篇文章