语步结构引导的科技论文结构化自动摘要方法研究*

张志豪　孙志春　张硕　林歌歌　梁国强
（北京工业大学经济与管理学院，北京 100124）

摘要：科技论文篇幅较长，各章节研究要点分散。现有自动摘要方法仅依据句子权重、段落位置筛选文本，易丢失关键信息，导致摘要逻辑断裂；通用大模型生成摘要虽然行文流畅，但存在输出结构难以约束、事实描述失真等问题。为此，本研究提出语步引导的科技论文结构化摘要模型——Move-driven Summarization（MoverSum）。MoverSum采用Seq2Seq多任务学习框架，以细粒度语步为全局规划线索，融合文本语义、章节位置、语步标签构建多维特征，同步完成句子重要度评估与摘要语步序列预测，按照语步序列抽取关键语句，以生成条理清晰的结构化摘要。本研究基于arXiv和PubMed数据集开展综合实验，结果表明：MoverSum的各项ROUGE指标全面优于各类基线模型；人工评估进一步证实，该模型生成摘要具备更完整的内容覆盖度与更流畅的行文逻辑。本研究经消融实验、大模型对比实验证实，多维特征与语步规划是模型性能提升的核心，MoverSum对科技文本摘要场景适配度更高。本研究提出方法兼顾抽取式摘要的事实保真度与输出结构可控性，为科技论文结构化摘要任务提供全新语步感知建模思路。

关键词：科技论文；结构化摘要；语步结构；Seq2Seq框架；多任务学习
中图分类号：G253　DOI：10.3772/j.issn.1673–2286.2026.05.004
引文格式：张志豪，孙志春，张硕，等. 语步结构引导的科技论文结构化自动摘要方法研究[J]. 数字图书馆论坛，2026，22（5）：34–47.

自动文本摘要旨在将长文档压缩为简洁的摘要，同时保留其核心信息，在缓解文本信息过载，帮助用户快速理解文本内容方面发挥着重要作用[1–4]。近年来，科技论文数量的爆炸式增长给科研人员文献检索、快速研读带来巨大负担，高效、精准的论文自动摘要技术日益受到关注。相较于新闻、社交媒体短文、推文等文本，科技论文通常专业性更强、篇幅更长、结构更复杂，对科技文本自动摘要方法提出了更高挑战。

现有科技论文自动摘要方法主要分为抽取式和生成式两类。抽取式方法通过选取并组合原文中的关键句子生成摘要，信息忠实、结果稳定[5]，但是可能出现缺乏语言连贯性等问题；生成式方法在理解原文语义的基础上重新组织信息，逐词生成原文中未直接出现的全新语句，容易出现语义混乱、事实错误等问题[6–7]。因此，在对准确性要求较高的科技论文自动摘要场景中，抽取式方法通常更具实用性。尽管此类方法的研究已取得一定进展，但忽视了科技论文中更细粒度的语步结构。语步是承载研究背景、方法、结论等核心表意

收稿日期：2026–03–06
*本研究得到国家自然科学基金委员会青年科学基金项目“大模型赋能的高价值专利个性化交易推荐方法与应用研究”（编号：72404020）、“颠覆性低碳技术识别及创新路径动态选择研究”（编号：72304023）及北京市自然科学基金面上项目“协同进化视角下研究前沿的涌现机制及识别方法研究”（编号：9232002）资助。

的修辞单元，是梳理文献逻辑的关键支撑。现有科技论文自动摘要方法缺少对语步信息的精细化建模，导致生成的摘要局限于局部信息，在逻辑连贯性、内容完整性及结构合理性等方面仍存在明显不足。

近年来，大语言模型（Large Language Model，LLM）在文本摘要、智能问答及学术文献理解与分析等任务中展现出卓越能力，推动了自动摘要研究范式的深刻转变。然而，LLM时代的到来并不意味着面向特定领域与任务的精细化自动摘要失去价值。恰恰相反，随着科技论文文档长度、结构复杂度与信息密度的持续攀升，如何实现对长篇科技论文的结构化理解、内容规划与可控生成，仍是当前自动摘要研究亟待解决的关键问题。

针对上述问题，本研究提出一种语步结构引导下基于Seq2Seq[8]（Sequence-to-Sequence）框架的科技论文结构化自动摘要方法，将细粒度语步作为显式的中间规划信号引入科技论文自动摘要任务。与现有研究采用“先识别、后抽取”的两阶段无监督范式不同，本研究实现了“语步预测—句子抽取”的联合优化，在生成过程中动态协调结构规划与内容选择，提升生成摘要的结构性、连贯性与信息覆盖度。此外，本研究构建融合句子语义、章节位置与语步标签的多层次输入表示机制，设计多任务学习框架，在统一架构中协同摘要的逻辑结构与内容选择，兼顾抽取式方法的信息忠实性与生成式方法的组织灵活性，为科技文本结构化自动摘要提供了一种高效、可控的新路径。

1　相关研究

鉴于抽取式摘要具备事实保真度高、可靠性更强的优势，且更适配科技论文摘要生成场景，本研究聚焦两类自动摘要生成方法：一是未使用语步结构的抽取式方法，侧重基于语义重要性选择关键句；二是引入语步结构的方法，强调利用句子的修辞功能引导摘要的内容选择与逻辑组织。在此基础上，本研究还关注LLM在该领域的最新进展，特别是其长上下文理解与结构化生成能力如何推动科技论文长文档摘要的研究。

1.1　科技论文抽取式摘要生成

科技论文抽取式摘要生成是通过识别并组合科技论文中的关键句子自动生成摘要，在很大程度上规避了失真问题，因而具有较高的内容可靠性。图排序算法是抽取式摘要的主流经典方法，核心原理为构建句子关联图，依托节点权重筛选核心语句。TextRank[9]与LexRank[10]是早期典型的无监督图排序抽取算法，二者借鉴PageRank[11]模型，以句子为图节点，通过句子相似度构建加权边、计算节点中心度，最终选取高得分句子组成摘要。后续研究多围绕图排序算法开展优化迭代：Zheng等[12]提出增强型图排序算法PacSum，引入文本相对位置因子优化节点中心度的计算方式；Dong等[13]与Zhu等[14]则构建“词—句—章节”多层异构图，依托章节边界剔除冗余关联信息，大幅提升了长文档摘要的生成精度。

近年来随着神经网络技术的发展，基于神经网络的抽取式摘要生成方法将摘要生成问题建模为句子分类和句子排序两类任务，比传统无监督方法的性能表现更好。句子分类方法通过对句子进行二分类实现摘要抽取，为每句分配“0”“1”标签，分别代表非摘要句与摘要句，最终将标签为“1”的句子拼接生成摘要。该方法的核心是学习优质的句子向量表征，并依托表征完成分类任务。在相关研究中，Nallapati等[15]与Cheng等[16]分别基于循环神经网络与卷积神经网络构建分层文档编码器，有效提升了抽取式摘要效果；Xiao等[17]则融合全局与局部语境信息，采用层次化长短期记忆网络挖掘文档全局特征与章节局部特征，缓解了长文本序列的建模难题。句子排序方法以句子重要性评分为核心，通过量化各句子作为摘要内容的置信度，对全文句子按分值降序排序，筛选Top-K高权重句子整合生成摘要。Zhou等[18]提出的NeuSum模型以句子增益为评分依据，兼顾句子间关联特征，可根据候选句对摘要整体的贡献动态筛选内容；Zhong等[19]提出的MatchSum模型将摘要抽取转化为文本匹配任务，保障了生成摘要与原文档的语义一致性。

尽管上述方法通过引入章节划分或图结构，在一定程度上增强了模型对科技文本结构的感知，但大多数现有研究将句子视为同质的语义单元，忽视了其在学术论证中所承担的细粒度修辞功能。由于缺乏对这类深层修辞意图的显式建模，现有方法难以生成逻辑缜密、功能完整的结构化摘要。

1.2　基于语步的科技论文结构化摘要生成

语步是学术论文中承担特定修辞功能的语句单元[20−21]，可实现阐述研究目的、说明研究方法、总结

研究结果等表达作用，其划分依据为句子在宏观话语结构中的修辞定位，是解析文献架构与核心内容的关键抓手。科技论文摘要具备规整的语步逻辑结构，以此为框架，可引导模型精准提取源文档核心内容，规范自动摘要的生成逻辑。目前学界已形成多种科技论文摘要语步分类体系，其中代表性方案为论辩分区（Argumentative Zoning，AZ）[22]与核心科学概念（CoreSC）[23]。AZ方案基于句子的论证意图，对文本背景、研究目的、研究方法等内容进行分类标注；CoreSC方案则从研究目标、方法、结果等维度划分论文的功能性单元。在此基础上，Cohan等[24-25]构建了科技文档句子级修辞标签的结构化分类与建模方案，实现句子修辞角色的自动分类。许钦亚等[26]结合ChatGPT数据增强技术与预训练语言模型，提出细粒度语步识别方法，有效提升了科技文本语步识别的精准度。上述研究推动了语步识别由基于规则和人工特征的方法向上下文感知的神经模型发展，为语步信息应用于自动摘要提供了较为可靠的结构化标签，但其研究重点仍主要是句子功能识别，尚未直接解决摘要内容选择与组织问题。

基于前人研究，Goldsack等[27]系统考查了科技文档中各章节的重要性及其对应摘要的语步分布，并将这一分析结果应用于自动摘要任务，提出一种融合领域知识与语步结构的两阶段抽取式摘要模型，该模型不仅架构简洁、性能优异，还具有较高的可解释性。在此基础上，习海旭等[28]提出了一种基于语步的科技论文结构化摘要生成框架：先按语步划分论文；然后在各语步内结合多特征加权TextRank抽取句子，利用MMR（Maximal Marginal Relevance）算法去除冗余，并通过依存句法分析精炼语义；最终融合各语步摘要生成结构清晰、内容凝练的摘要。上述研究采用“先识别、后抽取”两阶段无监督范式[27-28]，将语步识别与句子抽取割裂处理，缺乏端到端的联合优化机制，难以在生成过程中动态协调结构规划与内容选择。这种分离式方案不仅削弱了语步与关键信息之间的细粒度关联建模能力，还易引入级联误差，导致生成的摘要在逻辑连贯性、内容覆盖完整性及结构合理性等方面存在明显不足，制约了模型性能的进一步提升。

1.3 LLM应用于科技论文结构化摘要生成

传统模型受限于输入长度，处理长篇幅科技论文时常需分段截取，易导致跨章节关键信息丢失。而LLM上下文窗口的扩展，为长文本摘要提供了新的技术支撑。针对现有数据集难以还原长文本摘要难点的问题，Liu等[29]构建了面向综述摘要的SumSurvey数据集，弥补了文本长度与信息分布的局限。Li等[30]指出长文档信息分散，提出HERA框架，通过对文档片段进行语义重排序与上下文打包，提升了LLM的信息利用效率。Chen等[31]则融合分层切分与思维链策略，引导模型分层理解并组织摘要。尽管上述方法改善了长文档的信息利用，但对科技论文中背景、目标、方法等细粒度修辞功能的挖掘仍显不足。这表明，即便依托大模型，显式规划与结构化引导仍是提升长文本摘要质量的关键。

2 模型构建

2.1 构建思路

本研究提出一种语步结构驱动的结构化摘要生成模型——Move-driven Summarization（MoverSum），一方面以语步结构引导摘要内容筛选，另一方面借助Seq2Seq框架完成结构化摘要建模。科技论文摘要需遵循“背景—目标—方法—结果”的规范语步逻辑，并非关键句子的简单拼接。引入语步作为显式规划线索，可帮助模型识别各语句的论证功能，均衡覆盖各类关键信息，解决传统抽取式摘要结构零散、内容缺失的问题。此外，传统打分模型仅能判别句子重要性，无法建模语步间的逻辑依赖与排列顺序。Seq2Seq框架可弥补这一缺陷：其编码器融合语义、位置、语步多维特征，完成全文信息的全局建模；其解码器自适应生成标准摘要语步序列，为句子抽取提供结构约束。MoverSum依托多任务学习同步优化句子重要度预测与语步序列生成，在维持抽取式摘要事实保真优势的基础上，能显著改善摘要的结构完整性与逻辑连贯性。

MoverSum框架如图1所示，包括语步识别、多维表征融合和结构化摘要提取三个环节。首先，针对科技论文普遍缺乏细粒度语步标注的问题，本研究引入基于SciBERT的语步识别预处理模块。先使用标签正则化合并4个人工标注语步数据集，并基于SciBERT训练语步分类器；然后对输入文档中的每个句子进行自动分类，赋予其相应的语步标签（BAC、OBJ、APP、OUT、OTH），为后续建模提供功能语义层面的结构化

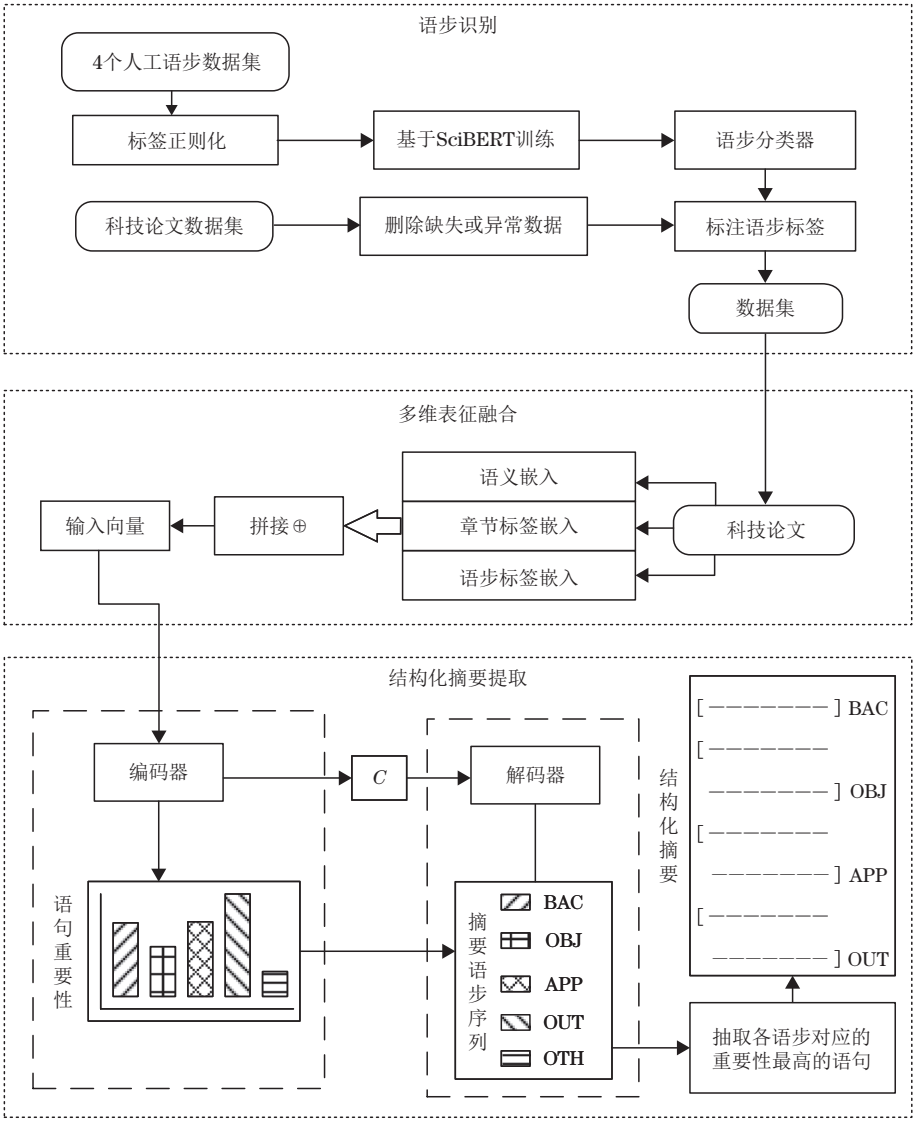


图1　MoverSum框架

先验知识。其次，构建融合语义、位置、功能三重信息的多层次输入表征。区别于传统仅依赖词向量的方法，MoverSum显式拼接三类嵌入，即语义嵌入（基于词向量的语义表征）、章节标签嵌入（编码句子在文档结构中的位置）以及语步标签嵌入（刻画句子在学术论证中的修辞功能）。3层次的向量拼接形成最终模型输入向量，该复合表征策略使模型在编码初期即可同步感知句子的内容语义、结构位置与话语功能，为后续的上下文建模奠定坚实基础。最后，引入“先规划—后抽取”的生成策略，模拟学者“构思提纲—填充内容”的写作过程完成结构化摘要提取。编码器对输入句子进行建模并输出其重要性评分，用于判断是否纳入摘要；解码器先生成目标摘要的语步序列，然后据此在科技论文原文中动态抽取各语步对应的重要性最高的语句，将这些句子有序组合，生成结构清晰、逻辑连贯的结构化摘要。

2.2　模型算法

MoverSum是一个基于多任务学习的Seq2Seq框架，包含编码器和解码器两个核心模块。①编码器：将文档中的句子按其原始顺序依次输入，为每个句子计算一个重要性分数，用于判断其是否应被纳入摘要。②解码器：基于编码器输出的上下文表示生成语步级的内容规划，并以此引导后续的句子选择与排序，从而生成结构合理、语义连贯的摘要。下文将详细介绍模型的整体架构与具体实现，并进一步阐述其训练策略，MoverSum算法原理如图2所示。

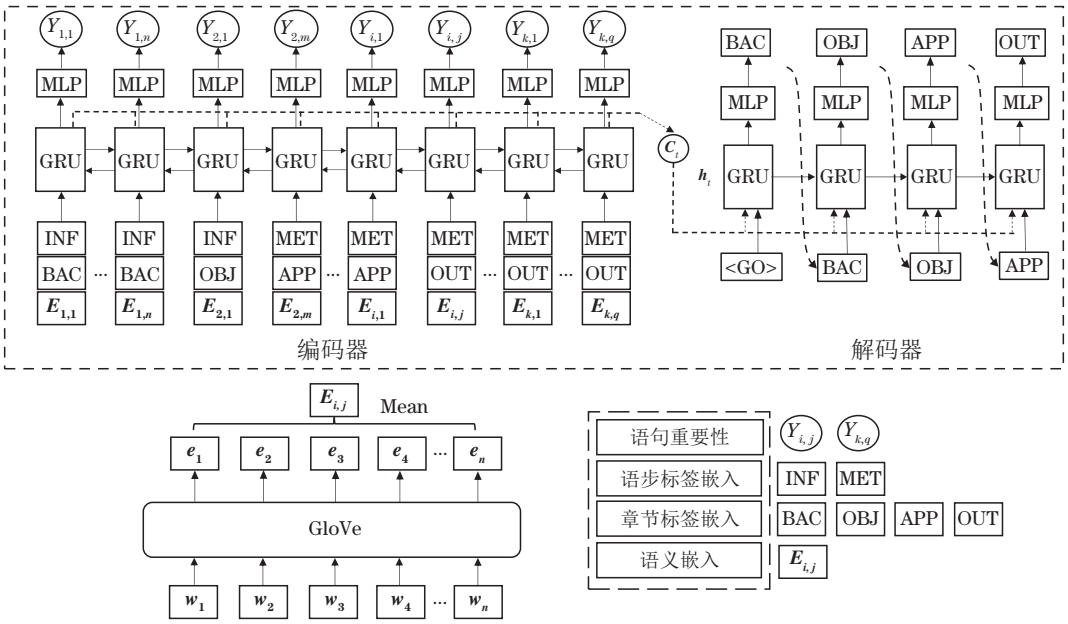


图2　MoverSum算法原理

2.2.1　编码器

编码器主要包括句子编码和文档编码两部分。在句子编码层面，给定包含多个句子的输入文档$D=\{s_{1,1}, s_{1,2}, \ldots, s_{2,1}, \ldots, s_{i,j}\}$，其中$s_{i,j}$表示文档中第$i$章节的第$j$个句子，每个句子由多个词汇组成，即$s_{i,j}=\{w_{i,1}, w_{i,2}, \ldots, w_{i,j}\}$，其中$w_{i,j}$表示句子$s_{i,j}$的第$j$个词。句子编码的目标是将输入的句子映射到一个固定长度的向量，并作为该句子的表征，如公式（1）所示。本研究采用词嵌入平均池化法来进行句子嵌入，句子的语义嵌入$\boldsymbol{S}_{i,j}$是其词嵌入的平均值，Emb(·)代表句子的语义表征模型。

$$\boldsymbol{S}_{i,j}=\frac{1}{j}\sum_{w_{i,1}}^{w_{i,j}}\mathrm{Emb}(w),\quad \boldsymbol{E}_{i,j}\in\mathbb{R}^{d_{\mathrm{emb}}} \tag{1}$$

考虑到不同的章节内容对最终摘要的贡献程度不同[27]，本研究引入句子$s_{i,j}$所属的章节标签嵌入$\boldsymbol{e}^{\mathrm{s}}_{i,j}$，以引导编码器关注信息量更高的段落，从而辅助其更准确地预测输入句子的重要性。此外，为进一步支持内容规划解码器生成精准的语步序列，还在编码器端引入句子$s_{i,j}$的语步标签嵌入$\boldsymbol{e}^{\mathrm{m}}_{i,j}$。如公式（2）所示，本研究将句子的语义嵌入、章节标签嵌入与语步标签嵌入三者拼接，构成输入句子$s_{i,j}$的最终表征$\boldsymbol{x}_{i,j}$。

$$\boldsymbol{x}_{i,j}=\boldsymbol{S}_{i,j}\oplus\boldsymbol{e}^{\mathrm{s}}_{i,j}\oplus\boldsymbol{e}^{\mathrm{m}}_{i,j} \tag{2}$$

其中，$\boldsymbol{e}^{\mathrm{s}}_{i,j}$是章节位置嵌入向量，s表示章节（section）；$\boldsymbol{e}^{\mathrm{m}}_{i,j}$是句子语步标签嵌入向量，m表示语步（move）；⊕表示向量拼接操作。

在文档编码层面，采用双向门控循环单元（BiGRU）[32–33]按对所有句子进行编码，如公式（3）所示。

$$\begin{cases}\overrightarrow{\boldsymbol{o}^{\mathrm{E}}_i},\overrightarrow{\boldsymbol{h}^{\mathrm{E}}_i}=\overrightarrow{\mathrm{GRU}^{\mathrm{E}}}\left(\boldsymbol{x}_i,\boldsymbol{h}_{i-1}\right)\\ \overleftarrow{\boldsymbol{o}^{\mathrm{E}}_i},\overleftarrow{\boldsymbol{h}^{\mathrm{E}}_i}=\overleftarrow{\mathrm{GRU}^{\mathrm{E}}}\left(\boldsymbol{x}_i,\boldsymbol{h}_{i+1}\right)\\ \boldsymbol{o}^{\mathrm{E}}_i=\overrightarrow{\boldsymbol{o}^{\mathrm{E}}_i}\oplus\overleftarrow{\boldsymbol{o}^{\mathrm{E}}_i}\end{cases} \tag{3}$$

假设全文一共有m个句子，给定输入序列$\boldsymbol{x}_{1:m}=[\boldsymbol{x}_1, \boldsymbol{x}_2,\ldots,\boldsymbol{x}_m]$。BiGRU输出对应的句子级上下文表征$\boldsymbol{o}^{\mathrm{E}}_{1:m}=[\boldsymbol{o}^{\mathrm{E}}_1, \boldsymbol{o}^{\mathrm{E}}_2, \ldots, \boldsymbol{o}^{\mathrm{E}}_m]$，其中上标E代表BiGRU编码器，方向箭头（→/←）分别代表前/后向的GRU编码器，$\boldsymbol{o}^{\mathrm{E}}_i$代表编码器的第$i$个输出，代表句子$i$在编码器端对应的嵌入表征。

如公式（4）所示，编码器输出的句子表征$\boldsymbol{o}^{\mathrm{E}}_{1:m}$被进一步输入至一个多层感知器（Multi-Layer Perceptron，MLP），后接激活函数（ReLU和Sigmoid），用于预测各句子的重要性得分$\boldsymbol{p}_i$。

$$\begin{cases}\boldsymbol{h}_i=\mathrm{Dropout}\left(\mathrm{RELU}\left(\boldsymbol{W}^1_{\mathrm{MLP}}\boldsymbol{o}^{\mathrm{E}}_i+\boldsymbol{b}_{\mathrm{MLP}}\right)\right)\\ \boldsymbol{p}_i=\sigma\left(\boldsymbol{W}^2_{\mathrm{MLP}}\boldsymbol{h}_i+\boldsymbol{b}_{\mathrm{h}}\right)\end{cases} \tag{4}$$

其中，$\boldsymbol{W}^1_{\mathrm{MLP}}$和$\boldsymbol{W}^2_{\mathrm{MLP}}$为可学习的权重矩阵，$\boldsymbol{b}_{\mathrm{MLP}}$和$\boldsymbol{b}_{\mathrm{h}}$为对应的可训练的偏置向量。

在时间步结束后，编码器会将最后一个隐藏态$\boldsymbol{h}_t$以及上下文向量$\boldsymbol{c}_t$传递给解码器。上下文向量$\boldsymbol{c}_t$计算如公式（5）所示。

$$\begin{cases}\boldsymbol{\beta}_i=\left[\boldsymbol{h}_{t-1}^{\mathrm{D}}\oplus\boldsymbol{o}_i^{\mathrm{E}}\right]\boldsymbol{w}\\ \boldsymbol{\alpha}_i=\mathrm{softmax}\left(\boldsymbol{\beta}\right)\\ \boldsymbol{c}_t=\sum_{i=1}^{n+1}\boldsymbol{\alpha}_i\times\boldsymbol{o}_i^{\mathrm{E}}\end{cases}\tag{5}$$

其中，$\boldsymbol{o}_i^{\mathrm{E}}$是编码器的第$i$个输出，代表句子$i$在编码器端对应的嵌入表征$\boldsymbol{\alpha}=[\boldsymbol{\alpha}_1,...,\boldsymbol{\alpha}_{m+1}]$、$\boldsymbol{\beta}=[\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_{m+1}]$，$\boldsymbol{\beta}_i$和$\boldsymbol{\alpha}_i$分别代表归一化前后的注意力权重向量，代表$\boldsymbol{w}$是一个可训练的向量。

2.2.2　解码器

解码器为输入的科技论文生成对应的摘要语步结构，该语步结构可以指导相应摘要生成，即完成摘要的内容规划。在每个时间步t，解码器都会生成一个语步标签，迭代后可生成一个语步标签序列，该序列就是摘要的语步结构。当所有有效的语步标签均被生成时，解码器停止。

单向门控循环单元GRU[32]被用作解码器的基本单元。$\boldsymbol{u}_t$是解码器在时间步t的输入，计算方法如公式（6）所示。

$$\boldsymbol{u}_t=\left[\boldsymbol{e}_t^{\mathrm{s}}\otimes\boldsymbol{c}_t\right]\boldsymbol{W}^{\mathrm{u}}\tag{6}$$

其中，$\boldsymbol{c}_t$是上下文向量，$\boldsymbol{W}^{\mathrm{u}}$是权重矩阵，$\boldsymbol{e}_t^{\mathrm{s}}$是在解码器时间步$t$–1预测的语步标签的表征。在时间步$t$，解码器的输出$\boldsymbol{o}_t^{\mathrm{D}}$和隐藏状态$\boldsymbol{h}_t^{\mathrm{D}}$计算如下公式（7）所示。

$$\boldsymbol{o}_t^{\mathrm{D}},\boldsymbol{h}_t^{\mathrm{D}}=\mathrm{GRU}\left(\boldsymbol{u}_t,\boldsymbol{h}_{t-1}^{\mathrm{D}}\right)\tag{7}$$

其中，$\boldsymbol{h}_{t-1}^{\mathrm{D}}$是解码器在前一个时间步的隐藏状态。如公式（8）所示，解码器输出$\boldsymbol{o}_t^{\mathrm{D}}$被输入另一个MLP层，最终经过softmax进行语步标签预测。

$$\boldsymbol{e}_t=\mathrm{softmax}\left(\boldsymbol{W}_{\mathrm{MLP}}^{\mathrm{d}}\boldsymbol{o}_t^{\mathrm{D}}+\boldsymbol{b}_{\mathrm{MLP}}^{\mathrm{d}}\right)\tag{8}$$

其中，$\boldsymbol{W}_{\mathrm{MLP}}^{\mathrm{d}}$和$\boldsymbol{b}_{\mathrm{MLP}}^{\mathrm{d}}$是MLP权重矩阵和偏置向量，$\boldsymbol{e}_t$是MLP输出的摘要语步预测。

2.2.3　模型训练

MoverSum是一个基于多任务学习的Seq2Seq框架。本研究的损失函数计算方法如公式（9）所示，采用批量梯度下降算法对其进行优化。

$$\mathcal{L}=\lambda\cdot\mathcal{L}^{\mathrm{E}}+\mathcal{L}^{\mathrm{D}}\tag{9}$$

其中，$\mathcal{L}$是模型总损失，$\mathcal{L}^{\mathrm{E}}$是编码器损失，$\mathcal{L}^{\mathrm{D}}$是解码器损失，λ是常数系数，用于平衡两个损失的权重系数。

3　实验过程

3.1　数据集构建与预处理

本研究选用arXiv和PubMed两个主流的大规模公开科技论文数据集作为实验语料[24]。其中，arXiv涵盖数学、物理学、计算机科学等多个学科，PubMed聚焦于生物医学领域，两个数据集规模较大，文档长度和摘要长度均明显高于新闻摘要等短文本数据集，更符合本研究对长篇科技论文摘要生成问题的研究设定。本研究的实验数据集统计情况如表1所示。

表1　实验数据集统计情况

数据集名称	文档数/篇	文档平均长度/词	摘要平均长度/词
PubMed	133 000	3 016	203
arXiv	215 000	4 938	220

为适配本研究的模型设计，对上述两个数据集进行以下标准化处理。

（1）抽取式标签生成。由于原始数据集提供论文正文及作者撰写的摘要，但未直接提供正文句子的抽取标签。为训练抽取式摘要模型，需要依据作者摘要构造句子级监督信号。具体而言，利用SciBERT[34]对句子进行语义编码，并计算其与人工撰写摘要语义的余弦相似度。选取相似度最高的若干原文句子作为最优抽取句（标记为“1”），其他句子则为非摘要句（标记为“0”）。

（2）章节标题标准化。科技论文通常遵循规范的篇章结构，包含引言、方法、实验结果等典型章节。章节标题的数量一般不超过10个，常见章节标题在形式、顺序和关键词使用上具有较强规律性。参照已有研究的章节归一化思路[27,35]，并结合样本中章节标题的实际分布，本研究采用正则表达式对标题字符串进行匹配，将符合相应模式的标题映射为统一的标准类别。章节标题标准化映射规则如表2所示，具体执行逻辑为：对章节标题字符串应用预设的正则表达式；若标题中出现与某一标准类别对应的匹配模式，则将该标题映射为相应标准类别。

需要注意的是，章节结构与语步结构是不同的概念，二者从不同粒度和维度刻画科技论文的结构。章节结构是论文显式的宏观组织单元，属于宏观位置线索。而语步结构是句子级的修辞功能结构，用于刻画

表2　章节标题标准化映射规则

标准化词	正则匹配模式
Introduction	introduction
Conclusion	conclu*; summary
Discussion	discussion
Result/Analysis	result; analys*; ablat*
Background/Motivation	background; motivation
Method	implement; method
Model	architect; system; model
Future Work	direction; future
References	referenc*
Acknowledgements	acknowledge
Related Work	related

注：其中“*”是正则通配符。

句子在学术论证中承担的功能，属于细粒度功能语义线索。

（3）语步标签增强与标准化。由于原始数据集缺乏细粒度的语步标注，本研究构建了一个基于SciBERT的语步分类器，用于对原数据集中科技论文缺失的语步信息进行自动语步标注。为训练该分类器，本研究选取ART Corpus、CSAbstract、AZ Article和AZ Abstract四个人工标注语步语料库。在统一各语料库的标签体系后，将其合并为训练数据。鉴于各语料库采用的标签体系存在差异，本研究设计了一套简明的标签映射规则，将不同标签统一归并为五类标准语步，即Background（研究背景）、Objective（研究目标）、Approach（研究方法）、Outcome（研究结果）和Other（其他），多源语步标签的标准化映射体系如表3所示。基于这一标准化后的合并数据集重新训练语步分类器，语步分类器的测试集结果如表4所示。语步分类器在上述合并的语步语料库中划分出的测试集上F1值为0.811、准确率为0.824，相比单一数据集略有提升，说明其能够较稳定地识别科技文本中的句子级修辞功能，可为后续语步感知摘要模型提供可靠的结构化输入。将其应用于arXiv与PubMed数据集中每篇论文的正文句子，为其自动分配语步标签，最终构建出适用于本研究的增强版标注数据集。

表3　多源语步标签的标准化映射体系

标准化标签	原始标签			
	ART Corpus	CSAbstract	AZ Article	AZ Abstract
BACkground	Background	Background	Background Connection Difference	Background Related Work
OBJective	Motivation Goal Hypothesis Object	Objective	Problem Future Work	Objective Future Work
APProach	Experiment Model Method	Method	Method	Method
OUTcome	Observation Result Conclusion	Result	Result Conclusion	Result Conclusion
OTHer		Other		

表4　语步分类器的测试集结果

训练数据	F1值	准确率
CSAbstruct	0.794	0.819
合并后数据集	0.811	0.824

3.2　基准模型和评估指标

为验证MoverSum的有效性，本研究选取了多类代表性的模型构建基线体系，其中，Oracle为抽取式摘要的理论性能上界，无监督抽取式基线模型包括Lead-3、PacSum[12]、HipoRank[13]、DODORANK[27]，有监督结构化基线模型选取Local & Global-Ext[17]和Longformer-Ext[17]。

在评估体系上，采用ROUGE自动评估[36]与多维度人工评价相结合的策略。自动评估使用ROUGE-1、ROUGE-2和ROUGE-L指标[37]（以下简称R-1、R-2和R-L），分别从单词、短语层面衡量生成摘要对原文信息的覆盖程度，并通过最长公共子序列反映其整体连贯

性与流畅性[16]。鉴于自动评估指标难以充分捕捉语义完整性与逻辑结构，本研究进一步引入三维度人工评价机制：覆盖率用来评估关键信息的完整性；重要性用来衡量内容对读者的核心价值；连贯性则考查摘要是否具备“背景—目的—方法—结果”等典型学术语步所构成的完整论证逻辑。

3.3　实验设置

本研究基于PyTorch框架搭建模型，词嵌入采用300维预训练GloVe向量[38]初始化，章节标签与语步标签均随机初始化为32维向量，以对结构特征进行建模；编码器选用隐藏层维度为300维的BiGRU，以有效捕捉上下文语义。训练过程中，设置Dropout概率为0.3，学习率设为1×10^{-4}，多任务损失函数中权重系数λ设为1；模型训练的batch size为64，采用随机梯度下降策略优化模型，并结合早停机制——当验证集损失在连续若干epoch内未下降时终止训练，保存最优模型权重。实验参数配置如表5所示。为降低随机因素对实验结果的影响，本研究通过设置不同的随机种子，对模型进行了5次独立重复实验，并以各项性能指标的平均值作为最终评估结果。

表5　实验参数配置

参数名称	参数值
词向量的维度	300
BiGRU的隐藏层	300
章节标签表征	32
句子修辞标签表征	32
损失权重系数λ	1
Dropout概率	0.3
学习率	1×10^{-4}
语步序列长度（arXiv）	8
语步序列长度（PubMed）	10

4　实验结果及分析

4.1　自动评估

MoverSum与选定基线方法在arXiv和PubMed两个数据集上的实验结果如表6所示，其中Oracle是在只能从原文挑句子规则下，与标准答案词汇重合度（ROUGE-1）最高的完美句子组合，即抽取式模型的（ROUGE-1）理论上限。

实验表明，本研究提出的MoverSum在两个数据集上的自动评估指标得分均优于所有对比基线模型。具体而言，在与无监督模型的对比中，引入细粒度语步信息的DODORANK显著优于未利用此类结构信息的其他基线，验证了语步建模在抽取式摘要任务中的有效性。相比之下，MoverSum通过多任务学习动态建模语步结构，实现了更精细、更具适应性的摘要生成。在与有监督基线的对比中，在arXiv和PubMed两个数据集上均优于Local & Global-Ext与Longformer-Ext。尤其是在PubMed数据集上，相较于Longformer-Ext，MoverSum在R-1、R-2和R-L上分别提升3.80、2.92和3.54个百分点，说明动态建模语步结构能够有效地支持长文档中的关键信息筛选与摘要生成。在与无监督基线的对比中，MoverSum在arXiv数据集上，相较于DODORANK模型在R-1、R-2和R-L指标上分别提升了5.66、3.41和5.69个百分点。

4.2　消融实验

为深入理解并系统评估MoverSum模型中关键组件对性能的贡献，本部分设计了三组消融实验，结果如表7所示。实验组1移除编码器端的多维表征融合模块，即移除章节标签与语步标签的嵌入表示，仅保留原始文本编码；实验组2移除解码器端的内容规划模块，即不生成摘要的语步结构，直接进行抽取；实验组3将解码器自动生成的语步序列替换为摘要对应的真实语步结构，以评估内容规划模块的生成质量对最终摘要性能的影响。上述实验可分别量化语步规划、结构感知表征以及规划准确性对模型整体效果的贡献。

4.3　敏感性分析

解码器生成的语步结构序列长度是MoverSum模型的重要超参数，它控制了模型在批训练过程中对序列的截断与填充长度，决定了生成摘要中语步序列的最大长度，从而直接影响模型生成摘要的质量。为探究该参数设置对模型性能的影响，需进行敏感性分析。

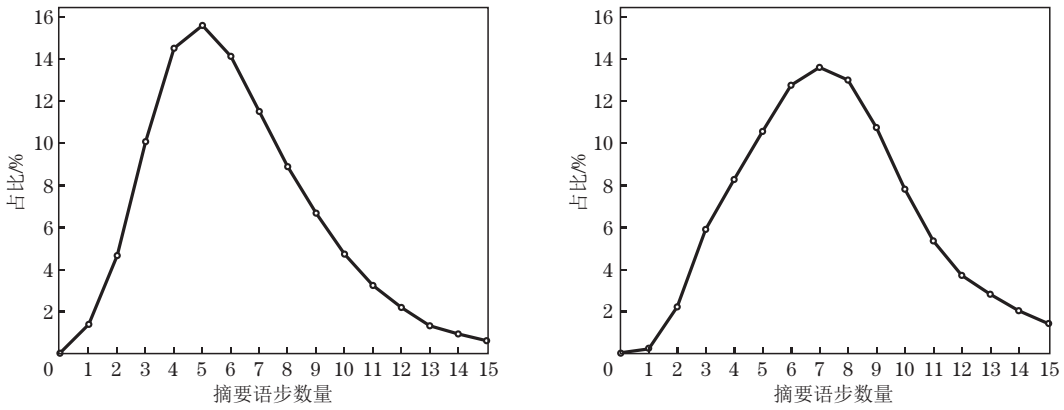
arXiv与PubMed数据集中真实摘要语步结构序列长度的分布情况如图3所示。可以看出，两个数据集的摘要语步结构序列长度分布均呈现出一定的集中趋势，大多数摘要的语步结构序列长度集中于6~10。为充分

表6　实验结果

组别	arXiv			PubMed		
	R-1	R-2	R-L	R-1	R-2	R-L
Oracle	53.88	23.05	34.90	55.05	27.48	38.66
Lead-3	33.66	8.94	22.19	35.63	12.28	25.17
PacSum	38.57	10.93	34.33	39.79	14.00	36.09
HipoRank	39.34	12.56	34.89	43.58	17.00	39.31
DODORANK	40.11	14.20	35.31	47.13	20.48	42.95
Local & Global-Ext	43.62	17.36	29.14	44.85	19.70	31.43
Longformer-Ext	45.24	16.88	40.03	43.75	17.37	39.71
MoverSum	45.77	17.61	41.00	47.55	20.29	43.25

表7　消融实验结果

实验组别	arXiv			PubMed		
	R-1	R-2	R-L	R-1	R-2	R-L
实验组1	44.58	17.57	40.04	46.66	19.62	42.35
实验组2	46.24	18.59	41.71	47.64	20.61	43.36
实验组3	46.68	17.84	41.74	47.85	20.42	44.90
MoverSum	45.77	17.61	41.00	47.55	20.29	43.25



(a)arXiv　　(b)PubMed

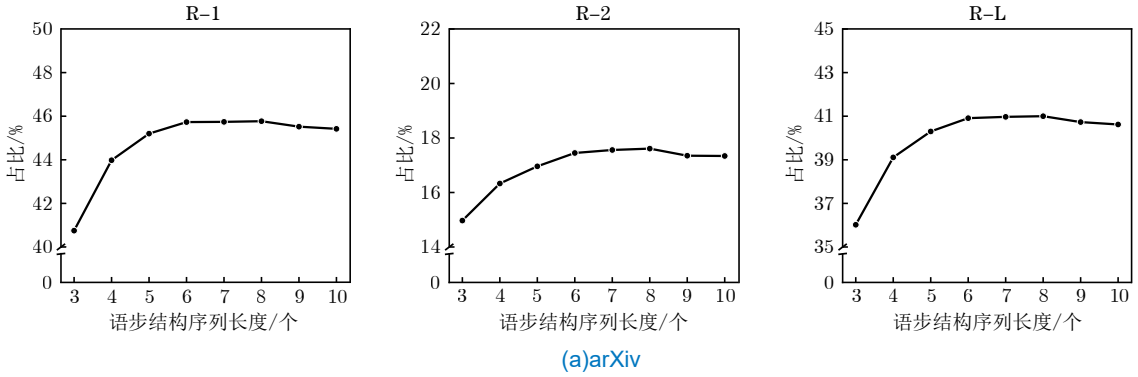
图3　arXiv与PubMed数据集中真实摘要语步结构序列长度的分布情况

覆盖主要分布区间，本研究将语步结构序列长度的搜索范围分别设为：arXiv数据集3~10，PubMed数据集5~12，步长均为1。

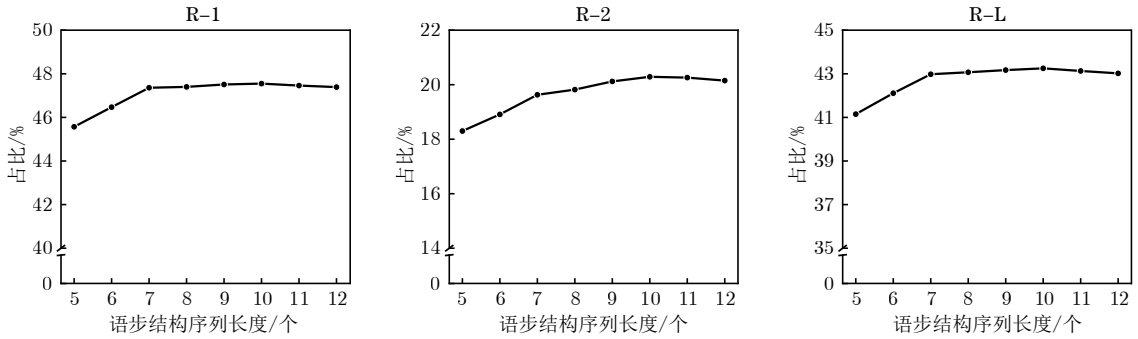
语步结构序列长度对arXiv与PubMed数据集的ROUGE指标影响如图4所示，模型性能随语步结构序列长度的变化呈现出“先上升后下降”的趋势。当长度设置过短时，摘要在训练过程中被强制截断，导致位于序列末端的关键语步被物理移除，阻碍了模型对完整学术论证逻辑的学习；而当长度设置过大时，则会引入冗余内容，降低摘要的简洁性与聚焦度，同样损害生成质量。实验表明，MoverSum在arXiv和PubMed数据集上分别于长度为8和10时取得最优性能，与两数据集中真实摘要语步数量的分布高度吻合。这一结果表明：只有当语步结构序列长度精准契合领域专家的自然写作习惯时，模型才能在避免关键信息截断与抑制无效噪声之间实现最佳平衡，从而充分释放结构化摘要生成的潜力。

4.4　人工评估

本部分选取Local & Global-Ext基线模型作为对照组。该模型仅基于Seq2Seq框架的编码器，通过注



(a)arXiv



(b)PubMed

图 4　语步结构序列长度对ROUGE指标的影响

意力机制融合输入文本的局部与全局信息，直接完成摘要句子的抽取。因此，它构成了验证本研究所提出的“语步预测—句子抽取”模式有效性的理想对比基准。为确保评估的客观性，本研究邀请两位具备相关领域背景的专家进行单盲评估，对从arXiv与PubMed测试集中随机抽取的50组样本进行独立评分。评估体系涵盖三个核心维度：内容覆盖率（衡量摘要是否完整涵盖原文关键信息）、要点重要性（评估所选内容的信息价值）、语义连贯性（考查摘要是否呈现出“背景—目的—方法—结果”等典型语步构成的完整学术论证逻辑）。所有维度均采用李克特量表[39]进行打分，取值范围为1~5分，1分为最低分，5分为最高分，得分越高表示样本质量越优。人工评估结果如表8所示。

表8　人工评估结果

评估维度	arXiv		PubMed	
	MoverSum	Local & Global-Ext	MoverSum	Local & Global-Ext
内容覆盖率	4.15	3.70	4.40	3.85
要点重要性	4.35	4.15	4.70	4.20
语义连贯性	3.40	3.26	4.10	3.45

评估结果显示，MoverSum在arXiv和PubMed数据集的三个维度上均优于Local & Global-Ext，尤其在语义连贯性方面优势显著，表明人类专家更偏好其生成的摘要。这一结果验证了本研究的核心假设：显式的“内容规划”模块作为逻辑蓝图，能将碎片化的抽取句子有机整合为结构清晰、连贯自然的摘要，有效缓解传统抽取式方法常见的逻辑断裂与拼凑感。此外，两位专家在arXiv和PubMed数据集上的Cohen’s Kappa系数[40]分别为0.51和0.58，表明尽管评估具有一定主观性，但在MoverSum的结构优势上达成稳健共识，进一步印证了评估结果的可靠性。

4.5　案例分析

为进一步验证生成摘要在语义连贯性方面的质量，本研究选取测试集中的典型样本进行案例分析。结果如图5所示，MoverSum与基线模型Local & Global-Ext虽均能识别关键信息，但在信息组织逻辑上存在明显差异。MoverSum生成的摘要基本遵循“研究背景—具体目标—核心发现”的科技文本修辞结构，体现出内容规划模块对解码过程的有效引导：在筛选高重要性句子的同时，兼顾其语步功能与位置合理性。相比之

MoveSumer

[Chronic Hepatitis B virus (HBV) infection is a major cause of hepatocellular carcinoma (HCC). | It has been reported that the HBx protein activates ERKs and PI-3K/Akt pathways [25], and further activates matrix metalloproteinase-9 [19] with a potential function of invasion and metastasis. | Interestingly, the activation of Rac1 and HBV replication has been found to be significantly decreased in cells transfected with a mutant HBV genome.]^{BAC} [Hepatitis B virus (HBV) can cause chronic and acute infection and remains a global health problem with a considerable morbidity and mortality.]^{APP} [We now report that the Rea1 GTPase was involved synergistically with HBx in the migration of HBV-replicating cells.]^{OBJ} [Findings in this report have provided an extension on the consequences of the Rac1 activation by HBV replication, in terms of cell motility. | To our knowledge, this is the first report to show the potential relationship between the proline-rich domain in HBx (in the context of HBV replication) and cell motility, and should shed new light on the understanding of the mechanism of HCC development.| Our findings can be further validated by a careful design of HBx knockdown in the context ofthe whole HBV genome, without affecting proper viral replication. |) Furthermore, the involvement of Rac1 can be supported by titration experiments in which thesame amount of the HBV genome is introduced into the cells with various amounts of Rac1 (e.g., activated orinactivated forms).]^{OUT}

Local&Global-Ext

[Interestingly, the activation of Rac1 and HBV replication has been found to be significantly decreased in cells transfected with a p-a mutant HBV genome.]^{BAC} [We now report that the Rac1 GTPase was involved synergistically with HBx in the migration of HBV-replicating cells. | Findings in this report have provided an extension on the consequences of the Rac1 activation by HBV replication, in terms of cell motility.]^{OUT} [To our knowledge, this is the first report to show the potential relationship between the proline-rich domain in HBx (in the context of HBV replication) and cell motility, and should shed new light on the understanding of the mechanism of HCC development.]^{OTH} [It has been reported that the HBx activates ERKs and PI-3K/Akt pathways[25], and further activates matrix metalloproteinase-9[19] with a potential function of invasion and metastasis. | We have recently demonstrated the same activation of ERK1/2 and Akt by HBx using the SH3 binding domain located in HBx.]^{BAC} [Processes such as cell adhesion, cell migration, cell polarity, spindle formation and locomotion were covered.]^{OTH} [Our data are in line with earlier reports on the potential invasion ability of HBx through indirect activation of matrix metalloproteinase (MMPs).]^{OUT}

图5　摘要生成案例分析

下，Local & Global-Ext虽捕获了关键语句，但排序缺乏逻辑连贯性。例如，部分背景性语句被置于摘要后半段，违背常规认知顺序，干扰了信息接收。该案例直观印证了定量结果：显式的语步结构规划机制有助于缓解长文本摘要中常见的逻辑松散问题，在保障信息覆盖的同时，使生成结果更符合用户阅读习惯与认知。

4.6　与基于LLM的摘要方法对比实验

为进一步验证MoverSum在LLM时代的任务价值，本研究补充了与主流大模型的对比实验。本研究选取DeepSeek-V3.2和GPT-4o作为对比模型，在arXiv和PubMed数据集上分别随机抽取100个样本进行实验，将原文按句子段落拼接后输入大模型，提示其生成简洁、忠于原文的摘要，并限制字数不超过200字。对比实验结果如表9所示，MoverSum在两个数据集上的R-1、R-2和R-L指标均优于所对比的模型，表明本研究所提方法在结构化信息压缩与关键内容保留方面仍具有显著优势。

该结果表明，在科技论文结构化抽取式摘要这一特定任务中，显式引入语步结构能有效提升摘要质量。需要强调的是，本实验并非旨在否定LLM在通用摘要

表9　对比实验结果

对比模型	arXiv			PubMed		
	R-1	R-2	R-L	R-1	R-2	R-L
DeepSeek-V3.2	44.06	14.75	39.57	43.66	14.55	39.00
GPT-4o	39.88	12.56	36.03	42.39	13.21	38.48
MoverSum	45.77	17.61	41.00	47.55	20.29	43.25

任务中的能力，而是进一步证实：针对科技论文这类结构严谨、信息密度高的长文档，语步结构建模能够为摘要生成提供有效的任务先验。本研究提出的“语步预测—句子抽取”模式，既可作为独立的抽取式摘要模型，也为后续研究提供了可迁移的框架。未来工作可将其融入LLM框架，用于摘要生成前的内容规划、语步级关键信息检索、候选句重排序，或作为生成结果的结构化校验模块。

5　结语

针对科技文本摘要任务中长文档结构复杂、信息分布稀疏以及生成结果逻辑松散等核心挑战，本研究提出一种语步结构引导的Seq2Seq结构化摘要生成模型MoverSum。该模型突破了传统抽取式方法仅依赖句子局部显著性的局限，形成了“先规划语步蓝图，后填充内容细节”的端到端生成模式。通过深度融合句子语义表征、章节位置信息与细粒度语步标签，MoverSum在保障关键信息覆盖度的同时，显著提升了生成摘要的逻辑连贯性与结构完整性。

在arXiv与PubMed数据集上的系统实验表明，MoverSum在ROUGE自动评估指标上显著优于多种基线模型，有效验证了本研究提出的“语步预测—句子抽取”模式的优越性；消融实验进一步证实，语步引导机制对性能提升具有关键作用。人工评估与敏感性分析也一致显示，该方法生成的摘要在结构完整性与逻辑流畅性方面优势显著，且其最优语步规划长度与专家的写作习惯高度吻合。本研究为科技论文结构化摘要生成提供了一种新颖、有效且可扩展的语步感知建模范式。

在LLM快速发展的背景下，未来研究可探索融合LLM与语步驱动的结构化摘要方案，持续提升科技文本摘要的智能生成能力、结构规范性与过程可解释性，为学术知识的高效传播与深度利用提供更有力支撑。

参考文献

[1] 夏吾吉，黄鹤鸣，更藏措毛，等. 基于无监督学习和监督学习的抽取式文本摘要综述[J]. 计算机应用，2024，44（4）：1035–1048.

[2] 李金鹏，张闯，陈小军，等. 自动文本摘要研究综述[J]. 计算机研究与发展，2021，58（1）：1–21.

[3] 其其日力格，斯琴图，王斯日古楞. 基于深度学习的自动文本摘要研究综述[J]. 计算机工程与应用，2025，61（18）：24–40.

[4] BELWAL R C，GUPTA A. Automatic text summarization techniques: a categorization, evolution and future scope[J]. Engineering Applications of Artificial Intelligence，2025，157：111216.

[5] ZHANG H，LIU X，ZHANG J. Extractive Summarization via ChatGPT for Faithful Summary Generation[C]//BOUAMOR H，PINO J，BALI K. Findings of the Association for Computational Linguistics：EMNLP 2023. Singapore：Association for Computational Linguistics，2023：3270–3278.

[6] CAO M，DONG Y，CHEUNG J. Hallucinated but Factual! Inspecting the Factuality of Hallucinations in Abstractive Summarization[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin：Association for Computational Linguistics，2022：3340–3354.

[7] LI T，LI Z，ZHANG Y. Improving Faithfulness of Large Language Models in Summarization via Sliding Generation and Self-Consistency[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics，Language Resources and Evaluation（LREC-COLING 2024）. Torino：ELRA and ICCL，2024：8804–8817.

[8] SUTSKEVER I，VINYALS O，LE Q V. Sequence to Sequence Learning with Neural Networks[C]//GHAHRAMANI Z，WELLING M，CORTES C，et al. Advances in Neural Information Processing Systems：Vol. 27. Curran Associates，Inc.，2014.

[9] MIHALCEA R，TARAU P. TextRank: Bringing Order into Text[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona：Association

for Computational Linguistics，2004：404–411.

[10] ERKAN G，RADEV D R. LexRank: graph-based lexical centrality as salience in text summarization[J]. J. Artif. Int. Res.，2004，22（1）：457–479.

[11] BRIN S，PAGE L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer Networks and ISDN Systems，1998，30（1）：107–117.

[12] ZHENG H，LAPATA M. Sentence Centrality Revisited for Unsupervised Summarization[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence：Association for Computational Linguistics，2019：6236–6247.

[13] DONG Y，MIRCEA A，CHEUNG J C K. Discourse-Aware Unsupervised Summarization for Long Scientific Documents[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics：Main Volume. Online：Association for Computational Linguistics，2021：1089–1102.

[14] ZHU T，HUA W，QU J，et al. Summarizing Long-Form Document with Rich Discourse Information[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. New York：Association for Computing Machinery，2021：2770–2779.

[15] NALLAPATI R，ZHAI F，ZHOU B. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents[C]//Proceedings of the AAAI conference on artificial intelligence. 2017.

[16] CHENG J，LAPATA M. Neural Summarization by Extracting Sentences and Words[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin：Association for Computational Linguistics，2016：484–494.

[17] XIAO W，CARENINI G. Extractive Summarization of Long Documents by Combining Global and Local Context[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）. Hong Kong：Association for Computational Linguistics，2019：3009–3019.

[18] ZHOU Q，YANG N，WEI F，et al. Neural Document Summarization by Jointly Learning to Score and Select Sentences[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne：Association for Computational Linguistics，2018：654–663.

[19] ZHONG M，LIU P，CHEN Y，et al. Extractive Summarization as Text Matching[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online：Association for Computational Linguistics，2020：6197–6208.

[20] SWALES J M. Research Genres: Explorations and Applications[M]. Cambridge：Cambridge University Press，2004.

[21] GUO Y，KORHONEN A，LIAKATA M，et al. Identifying the Information Structure of Scientific Abstracts: An Investigation of Three Different Schemes[C]//Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. Uppsala：Association for Computational Linguistics，2010：99–107.

[22] TEUFEL S. Argumentative zoning: Information extraction from scientific text[D].Edinburgh：University of Edinburgh，1999.

[23] LIAKATA M. Zones of conceptualisation in scientific papers: a window to negative and speculative statements[C]//Proceedings of the Workshop on Negation and Speculation in Natural Language Processing. Uppsala：University of Antwerp，2010：1–4.

[24] COHAN A，DERNONCOURT F，KIM D S，et al. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. New Orleans：Association for Computational Linguistics，2018：615–621.

[25] COHAN A，BELTAGY I，KING D，et al. Pretrained Language Models for Sequential Sentence Classification[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）. Hong Kong：Association for Computational Linguistics，2019：3691–3697.

[26] 许钦亚，薛秋红，钱力，等. 融合ChatGPT数据增强的学术论文语步识别方法研究[J]. 图书情报工作，2024，68（17）：84–94.

[27] GOLDSACK T，ZHANG Z，LIN C，et al. Domain-driven and discourse-guided scientific summarisation[C]//European Conference on Information Retrieval. Cham：Springer Nature Switzerland，2023：361–376.

[28] 习海旭，何胜，黄纯国. 融合语步和文本多特征的科技论文结构化摘要生成[J]. 情报学报，2023，42（10）：1176–1186.

[29] LIU R，LIU M，YU M，et al. SumSurvey: An Abstractive Dataset of Scientific Survey Papers for Long Document Summarization[C]//Findings of the Association for Computational Linguistics：ACL 2024. Bangkok：Association for Computational Linguistics，2024：9632–9651.

[30] LI T，CHEN H，YU F，et al. HERA: Improving Long Document

Summarization using Large Language Models with Context Packaging and Reordering[EB/OL]. [2026-05-02]. https://arxiv.org/abs/2502.00448.

[31] CHEN X，CHEN Z，CHENG S. CoTHSSum: Structured long-document summarization via chain-of-thought reasoning and hierarchical segmentation[J]. Journal of King Saud University Computer and Information Sciences，2025，37（4）：40.

[32] CHO K，VAN MERRIËNBOER B，GULCEHRE C，et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing（EMNLP）. Doha：Association for Computational Linguistics，2014：1724–1734.

[33] SCHUSTER M，PALIWAL K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing，1997，45（11）：2673–2681.

[34] BELTAGY I，LO K，COHAN A. SciBERT: A Pretrained Language Model for Scientific Text[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）. Hong Kong：Association for Computational Linguistics，2019：3613–3618.

[35] ERMAKOVA L，BORDIGNON F，TURENNE N，et al. Is the abstract a mere teaser evaluating generosity of article abstracts in the environmental sciences[J]. Frontiers in Research Metrics and Analytics，2018，3：16.

[36] LIN C Y. ROUGE: A Package for Automatic Evaluation of Summaries[C]//Text Summarization Branches Out. Barcelona，Spain：Association for Computational Linguistics，2004：74–81.

[37] KOH H Y，JU J，LIU M，et al. An Empirical Survey on long document summarization: datasets，models，and metrics[J]. ACM Computing Surveys，2023，55（8）：154.

[38] PENNINGTON J，SOCHER R，MANNING C. Glove: Global Vectors for Word Representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing（EMNLP）. Doha：Association for Computational Linguistics，2014：1532–1543.

[39] AMIDEI J，PIWEK P，WILLIS A. The use of rating and Likert Scales in natural language generation human evaluation tasks: A review and some recommendations[C]//Proceedings of the 12th International Conference on Natural Language Generation. Tokyo：Association for Computational Linguistics，2019：397–402.

[40] COHEN J. A Coefficient of agreement for nominal scales[J]. Educational and Psychological Measurement，1960，20（1）：37.

作者简介

张志豪，男，博士，讲师，研究方向：数据挖掘与自然语言处理。
孙志春，男，硕士研究生，研究方向：数据科学。
张硕，女，博士，副教授，研究方向：创新管理。
林歌歌，女，博士，讲师，研究方向：科学计量学。
梁国强，男，博士，副教授，通信作者，研究方向：创新管理与数据科学，E-mail：lianggq@bjut.edu.cn。

Research on Move-Guided Structured Automatic Summarization for Scientific Papers

ZHANG ZhiHao　SUN ZhiChun　ZHANG Shuo　LIN GeGe　LIANG GuoQiang
(College of Economics and Management, Beijing University of Technology, Beijing 100124, P. R. China)

Abstract: Scientific papers are typically lengthy, with key research points dispersed across various sections. Existing automatic summarization methods, which primarily rely on sentence weights and paragraph positions for text selection, often lose critical information, resulting in fragmented logic. While general-purpose large language models generate fluent summaries, they struggle to constrain output structures and frequently suffer from factual hallucinations. To address these challenges, this study proposes Move-driven Summarization (MoverSum), a move-guided structured summarization model for scientific papers. Built upon a multi-task learning Seq2Seq framework, MoverSum utilizes fine-grained rhetorical moves as a global planning guide. By integrating text semantics, section positions, and move labels into multi-dimensional features, the model simultaneously evaluates sentence salience and predicts the summary move sequence, extracting key sentences in accordance with the move sequence to generate well-organized, structured summaries. Comprehensive experiments conducted on the arXiv and PubMed datasets demonstrate that MoverSum consistently outperforms various baseline models across all ROUGE metrics. Furthermore, manual evaluation confirms that the generated summaries achieve superior content coverage and logical coherence. Ablation studies and comparative experiments with large language models verify that the multi-dimensional features and move-guided planning are the core drivers of performance improvements, highlighting MoverSum's high adaptability to scientific text summarization tasks. Ultimately, this proposed method balances the factual fidelity of extractive summarization with structural controllability, offering a novel move-aware modeling approach for the structured summarization of scientific papers.

Keywords: Scientific Paper; Structured Summaries; Move Structure; Seq2Seq Framework; Multi-task Learning