情报学报　2021 年 1 月　第 40 卷　第 1 期
Journal of the China Society for Scientific and Technical Information, Jan. 2021, 40(1): 43-52

DOI: 10.3772/j.issn.1000-0135.2021.01.005

学术文本词汇功能识别——基于标题生成策略和注意力机制的问题方法抽取

程齐凯[1,2]，李鹏程[1,2]，张国标[1,2]，陆　伟[1,2]

（1. 武汉大学信息管理学院，武汉　430072；2. 武汉大学信息检索与知识挖掘研究所，武汉　430072）

摘　要　学术文本词汇功能识别的目的是实现学术文本中表征问题、方法和对象等词汇的抽取。针对传统识别方法中训练难以获取所导致的识别准确率低、召回率有限和泛化能力差等问题，本研究提出了一种基于深度学习和标题生成策略的学术文本词汇功能识别方法，将任务形式由信息抽取转化为特定形式的标题生成问题。本研究采用构建seq2seq模型和引入注意力机制的方式捕获词汇多层语义信息，最终实现学术文本中问题和方法指代词的生成和获取。实验结果表明，通过应用深度学习方法和标题生成策略，本研究提出的模型能够从摘要中有效识别学术文献的主要研究问题和主要研究方法，并较已有方法在识别效果上有明显提升。

关键词　词汇功能识别；深度学习；自动文摘；学术文本

Recognition of Lexical Functions in Academic Texts: Problem Method Extraction Based on Title Generation Strategy and Attention Mechanism

Cheng Qikai[1,2], Li Pengcheng[1,2], Zhang Guobiao[1,2] and Lu Wei[1,2]

(1. *School of Information Management, Wuhan University, Wuhan*　430072;
2. *Institute for Information Retrieval and Knowledge Mining, Wuhan University, Wuhan*　430072)

Abstract: The purpose of academic text problem and method identification is to extract research questions and methods from academic text. Aimed at solving the problems of low recognition accuracy, limited recall rate, and poor generalization ability caused by the difficulty of obtaining the training set in traditional recognition methods, this study proposes an academic text problem recognition method based on a deep learning and title generation strategy. The method converts the extraction and recognition of the problem method into the form of title generation in a specific form. By constructing a seq2seq model and introducing an attention mechanism, multi-layer semantic word information was captured to generate and obtain the problem and method pronouns in academic texts. The experimental results showed that through the application of deep learning methods and title generation strategies, this study effectively identified core research problems and core research methods in academic literature.

Key words: lexical function recognition; deep learning; automatic abstraction; academic text

收稿日期：2020-05-16；**修回日期**：2020-10-11

基金项目：国家自然科学基金项目“基于多语义信息融合的学术文献引文推荐研究”（71673211）；国家自然科学基金青年科学基金项目“基于深度语义挖掘的引文推荐多样化研究”（71704137）。

作者简介：程齐凯，男，1989年生，博士，副教授，主要研究方向为自然语言处理、信息检索、机器学习；李鹏程，男，1994年生，博士研究生，研究方向为文本挖掘，深度学习；张国标，男，1990年生，博士研究生，研究方向为图像识别，深度学习；陆伟，男，1974年生，博士，教授，博士生导师，主要研究方向为信息检索、知识管理、数据智能等，E-mail：weilu@whu.edu.cn。

1 引 言

学术文本作为一种高信息密度的文档资源，是科研工作者实现知识生产和知识组织的重要载体。随着可获取数字图书资源的日益激增，“信息爆炸”和“信息过载”使得信息精准检索和知识快速获取越发困难[1]。即便是在面对一个相对较小的研究课题时，研究者也需要耗费大量时间和精力来完成相关文献的查阅工作。为方便研究者索引文献和获取知识，现有的符号系统制定了类目繁多的分类标引框架[2]，研究者通过使用统一普适的分类号来提高检索效率。然而，以文献为粒度单元的检索策略，并不能满足研究者逐渐增长的细粒度、导向性的知识快速获取需求。Ribaupierre 等[3]指出，科研人员信息获取行为往往基于目标和任务驱动，对于文章中的问题、方法或结果等特定语篇内容更为关注。因此，学者们试图在理解文本语义信息的基础上实现词汇粒度层面的文本标签构建，为知识密集型领域的知识服务体系提供底层索引支持。

学术文本词汇功能识别的目的是抽取出学术文本中表征的问题、方法、对象和工具等词汇，其本质为信息抽取问题。命名实体识别（named entity recognition，NER）作为信息抽取（information extraction，IE）领域中的重要下游分支，其任务形式与学术文本词汇功能识别也较为相似。鉴于命名实体识别的相关基础技术（如分词、词性标注、句法分析）都日趋完善，一种行之有效的策略是使用命名实体识别中的序列标注完成学术文本词汇功能的自动识别[11,18]。事实上，随着基于统计学习的有监督模型蓬勃发展，现有研究多将信息抽取问题转换为机器可解的标签判定问题或分类问题[4-5]，如在词汇功能识别任务中是判别每一个词汇或词汇组合是否属于特定类别。然而，“人工标注语料+机器学习算法”模式下的信息抽取需要大规模、高质量的标注语料来完成有监督学习模型的训练拟合，难以批量获取的源数据以及复杂烦琐的数据预处理，使得语料构建的成本不断攀升，由此造成现有判别式识别方法在准确率、召回率的提升上颇受掣肘。

在此背景下，本文提出了一种基于深度学习和标题生成策略的学术文本问题方法识别模型，应用 Encoder-Decoder 架构模型读取文本的语义特征，以自动文摘的任务形式生成能够揭示文本中核心问题与核心方法的特定样式标题，最终利用正则化实现问题方法的指代词汇抽取。相对于传统的词汇功能识别，本文所提出方法将功能性词汇的抽取识别转化为特定形式的标题生成问题，具有以下优点：①可直接利用数据库中所存有的大量规则样式标题作为模型的训练标签，省去了最为耗时费力的标注工作，使得高质量、大规模的语料构建成为可能；②本文能够从涉及多方法、多问题的学术文本中直接识别出具有对应关系的核心问题与核心方法，可为问题方法对应的知识库构建提供支持；③相比于判别式分类和序列标注的任务形式，序列到序列的功能词汇生成须在深层分析和理解文本语义的基础上实现，与人类行为模式更为契合。

全文后续内容安排如下：第 2 节简要介绍本文的相关研究现状，第 3 节详细描述基于标题生成策略的词汇功能识别模型构建，第 4 节为具体的实验过程以及实验结果，第 5 节在全文的基础上给出了总结。

2 相关工作概述

2.1 词汇功能识别

在自然语言处理领域中，学者们通常从语法、语义和语用三个层面对语言进行建模。语法研究是通过对语言结构的表示来描述语言符号的支配规则，早期的自然语言处理研究也多集中于此[6-8]，如分析句子主谓宾结构和词汇间依存关系的句法分析便是经典任务之一。在过去的二十余年里，语法层面的自然语言处理研究取得了较大发展，相关技术在诸多领域中也被广为应用[9-10]。随着统计学习和表示学习兴起，如何在语义和语用层面表征语句的字符含义以及理解当前语境下所表达的内容信息，成为了学者们的关注热点。

词汇是语言构成中最小的基本语义单元，词汇功能识别的目的则是从语义和语用的角度探究词汇在文本中所承载的功能角色[18]。Kondo 等[11]于 2009 年使用 CRF（conditional random field）模型对科技文献标题中的词汇进行“领域（head）”“目标（goal）”“方法（method）”及“其他（other）”的类别判定，根据得到的方法/技术来描绘特定领域内技术的演化路径和发展趋势。随后 Nanba 等[12]进一步将研究点聚焦于“技术（technology）”识别，应用 SVM（support vector machine）方法在专利文本上取得了 0.431 的召回率和 0.545 的准确率。针对专利分析，Trappey 等[13]及 Choi 等[14]使用“技术-功效”矩阵[15]实现专利文本中前沿技术的识别

挖掘。Gupta 等[16]使用句法模板从科技文献中识别出“话题（focus）”“技术（technique）”及“应用（application）”。在前者基础上，Tsai 等[17]对 Bootstrapping 算法进行了改进，使得计算量降低的同时提升了准确度。程齐凯[18]在已有文献的基础上对词汇功能的概念进行了界定，词汇或术语在文本中所承担角色，并构建了较为完善的学术文本词汇功能框架。此后，李信等[19]从语义理解的角度出发，依据程齐凯[18]所构建的词汇功能框架设计和实现了一个基于词汇功能识别的科研文献分析系统。刘智锋等[20]将词汇功能研究的判别对象限定为关键词，制定了计量学领域关键词语义功能分类框架：领域、对象、主题、方法和数据，并基于该框架构建了关键词语义功能标注数据集。

总而言之，词汇功能识别的相关研究仍处于初步探索阶段，出于研究目的和功能定义等主观因素，学者们并未能够就词汇的具体功能类别划分达成一致。除此之外，客观上存在的诸多制约也使得词汇功能的统一显得殊为不易。例如，每个学科或领域中均可能存在独有所属的功能类别，穷尽各个领域中的所有类别需要极大的工作量；再者，明确各个功能类别的划分界线，以及发现各个类别间的潜在上下位关系，也显得极其困难。通过对上述研究的梳理分析发现，尽管学者们在词汇功能类别的具体划分上不尽相同，但对于“问题”和“方法”的功能类别却表现出了一致的认同性。这是由于“问题驱动”在科学的进步乃至研究工作的推进中均扮演了关键角色。因此，本文沿用程齐凯[18]所提出的词汇功能划分体系，将学术文本词汇功能分为领域无关词汇功能和领域相关词汇功能。其中，领域无关词汇功能仅包含两类：问题和方法。研究问题与研究方法作为科技文献的核心知识单元，本文将聚焦于学术文本中领域无关词汇功能——研究方法和研究问题识别，通过采取标题生成策略和引入注意力机制的方法实现学术文本问题方法的指代词获取。

2.2　标题生成及作用机理

标题生成，是指用限定长度的单句对既定的信息内容进行概括表示，信息对象包括且不限于文本[21-22]、图像[23]以及视频[24-25]等。学术文本的标题生成可理解为全文层面的自动文摘任务，即将全文信息高度凝练为一定形式的规则短句，使得其能够扼要表示文本的核心研究内容。依据生成策略，自动文摘可分为抽取式和生成式两种。抽取式是对文档中的词或句进行重要性排序[26]，生成式则是在理解文本语义的基础上实现对原文的复述[27]。针对句子级层面的文本摘要任务，Nallapati 等[28]与 Ayana 等[29]分别使用抽取式和生成式方法进行了探讨。随着序列语言模型和 NLP 技术的日趋成熟，生成式文摘在语句可读性和关键信息完整性上得到显著提升，seq2seq+attention 组合方案也逐步成为生成式文摘中的经典模式[30-31]。鉴于生成式文摘的思想和过程与人类的行为模式更为贴近，本文采用基于 seq2seq 架构的生成式模型实现学术文本的标题生成，并引入注意力机制以优化标题的生成效果。

标题作为一篇文献的概括性描述，具有表达作者写作意图及文本主旨核心的重要作用。如 Hoey 所述，任何语篇中的阅读和写作过程都可看作是作者和读者之间一种交流互动，标题为该互动提供了一种可视化对话窗口[32]。Paiva 等[33]与 Jamali 等[34]的研究指出，标题中涵盖研究问题或研究结果的文献倾向于得到更高的阅览量和被引量。这是由于在现存形式的文献检索中，系统所返回的查询结果多表现为相关文献的标题罗列展示，其中读者试图通过标题信息预见作者将要回答的问题。在这一作用机理下，将研究问题和研究方法信息列入标题中，以直观揭示本文主旨核心的做法在当前并不鲜见。在现有的期刊数据库中，存在大量标题样式为“Research of A based on B”或“基于 A 的 B 研究”的期刊论文。考虑到这种规则特征在某种意义上是对文本研究内容的映射，Kondo 等[11]利用该思想在英、日文献的标题上实现了“领域（head）”“目标（goal）”和“方法（method）”等功能性词汇的抽取。此外，采用标题生成的方式对学术文本中的关键信息予以揭示的研究也不乏先例。例如，程齐凯[18]阐述了标题生成策略在文档级词汇功能揭示中的作用机理；Putra 等[35]则提出了一种涵盖文本研究目的（research purpose）及研究方法（research method）的标题生成模型，以供作者在拟定标题时作为备选参考。

本文借助标题生成的思想来完成问题方法描述词汇的获取。值得注意的是，Putra 等[35]与本文的任务目标较为相似，但在具体实验方法上与本文有较大区别，其在数据预处理中需将句子进行目标（AIM）、方法（OWN_MTHD）和其他（NR）的类别标注，本文参考程齐凯[18]的策略，利用现存有的规则标题直接完成 seq2seq 模型输出标签的获取。

3 研究方法

为了解决学术文本中词汇功能的自动识别问题，本文提出了一种基于标题生成策略的神经网络模型，通过将文本摘要转化成规则标题的形式，实现学术文本中研究问题与研究方法指代词的获取。简而言之，本文的研究任务可定义为：给定一个长度为 m 的文本序列 $T=\{s_1,s_2,\cdots,s_m\}$，生成长度为 n 的句子序列 $S=\{w_1,w_2,\cdots,w_n\}$（$m \gg n$），最终从序列 S 中抽取出所需的问题字符串 w_i 和方法字符串 w_j。

整体研究过程如下：①数据获取及预处理。包括数据的采集、清洗及标注等工作；②标题生成模型构建。采用基于 Encoder-Decoder 架构的语言序列模型，在理解文本语义的基础上，实现输入文本序列的摘要化，生成既定形式的规则标题；③问题方法指代词抽取。通过标题进行分词、词性标注及句法分析，利用规则匹配从中抽取出能够描述文本核心研究方法与核心研究问题的功能性词汇。

3.1 数据获取及预处理

现有基于有监督学习的词汇功能识别，偏好于采用分类或序列标注的方法来完成词汇的功能类别判定[17-18]，即通过在标注数据集上进行有监督训练，以实现问题方法等标签的功能判定。这一策略的缺点：必须为学习算法准备一定数量的高质量训练数据集，要求能够准确、完备标注出科技文献中问题方法等功能性词汇。同时，对于涉及多问题方法的文献，还需考虑该问题和方法究竟是作为文中的主要研究对象出现，还是仅仅作为参考背景而提及。

大规模的科学文献问题方法标注数据并不容易获取。首先，现实场景中的开放式陈述使得研究方法和研究问题具有诸多变体和表达形式；其次，需要在对文献仔细分析的基础上才能完成核心问题与核心方法的标注工作；最后，必须有多名行业专家参与数据标注，以避免文档主题的单一性。为克服训练数据的获取难题，文本提出了一种将信息抽取问题转化为标题生成问题的词汇功能识别方法，从待识别的全文或者摘要中生成类似于“基于[方法]的[问题]”的样式标题，继而间接识别出能够描述学术文献中核心问题和核心方法的功能性词汇。

在中文学术文本中，存在着大量的类似于“基于X的Y研究”样式的标题。与此相对应地，ACL数据库和ACM数据库收录的论文中也存在着大量类似“X based on Y”“X using Y”“Y algorithm based X”的样式标题。这些标题在一定程度上明确揭示了论文的核心问题和核心方法。图1给出了一个标题与摘要的标注示例。在所示论文中，标题的形式为“基于X的Y方法”，标题文本给出了该文档的核心问题和核心方法，分别是“微博情感分类”和“监督学习”，这些词汇或者词汇的同近义词也同时在摘要中出现。

基于上述分析，本文将核心问题和核心方法的识别问题转化为利用摘要（或全文）生成“基于X的Y研究”这一标题的问题。相对于前一问题，后一问题更容易解决，且后者的训练数据更容易获得。在学术数据库中，存在着大量标题形似“基于X的Y”的论文，这些论文的标题和对应的摘要（或全文）构成了标题生成模型训练天然存在的标注数据。

3.2 基于Encoder-Decoder的标题生成模型描述

Encoder-Decoder是seq2seq模型中的一种经典架构，其由三个部分组成：Encoder、Decoder以及连接两者的中间状态向量。其中，Encoder模块负责对输入信号的特征读取，将所输入的文本序列编码

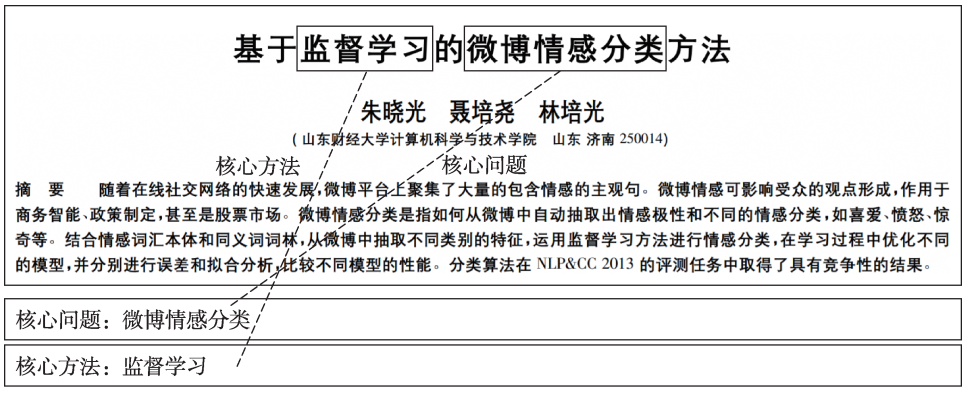

图1 论文标题与摘要的对照示例

成一个固定大小的状态向量 $\boldsymbol{W}$。待 Encoder 逐步完成输入的编码操作后，将包含全部特征信息的 $\boldsymbol{W}$ 传给 Decoder，再通过 Decoder 对状态向量 $\boldsymbol{W}$ 的学习来进行输出。

在图 2 所示的经典 Encoder-Decoder 模型结构图中，每一个 box 代表了一个语义读取单元（通常是 LSTM（long short-term memory）或者 GRU（gated recurrent unit）），用以捕获输入序列的语义信息。待得到包含序列语义特征的中间状态向量 $\boldsymbol{W}=F(A,B,C)$ 后，由 Decoder 模块对 $\boldsymbol{W}$ 进行解码操作，在每个时间步生成当前状态的语义输出 X、Y、Z，其中，$X=f(\boldsymbol{W}),Y=f(\boldsymbol{W},X),Z=f(\boldsymbol{W},X,Y)$。

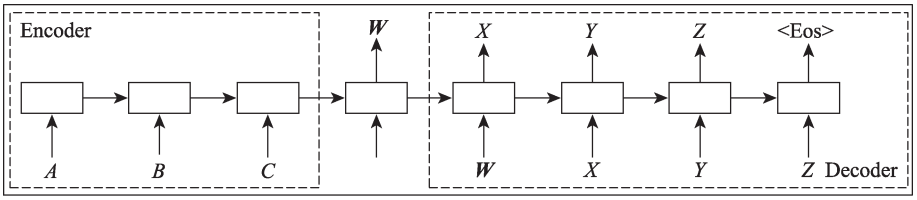

图 2　Encoder-Decoder 模型

学术文本的标题自动生成是在 Encoder-Decoder 架构基础上完成，具体模型如图 3 所示。输入层为预处理过的学术文本序列，对于每一条摘要为 $\{S_1,S_2,\cdots,S_m\}$ 的数据语料，均对应标签为{基于…的…}样式的规则标题；在嵌入层中，使用 word2vec 方法[36]对输入层中文本进行向量化表征，完成字符转向量的操作；随后，将该特征向量传至由双向 LSTM 所构成的 Encoder 层中并实现输入信息的语义编码，通过 LSTM 中的前后向迭代捕获文本中的潜在语义信息。此外，为力求文献摘要与生成标题间信息的充分交互，本文在编码层与解码层间引入了注意力机制，该机制可有效解决生成式文摘中的信息冗余问题，并广泛应用于 seq2seq 架构神经网络模型中[30]。本文通过使用注意力机制学习不同词位在标题生成中的权重信息，以减少因文本字符长度增加而造成的细节丢失。最终，由同样是双向 LSTM 所构成的 Decoder 层对中间层向量进行语义解码，并在全连接层输出能够揭示文中研究问题与研究方法的规则样式标题——基于 XX 的 XX。

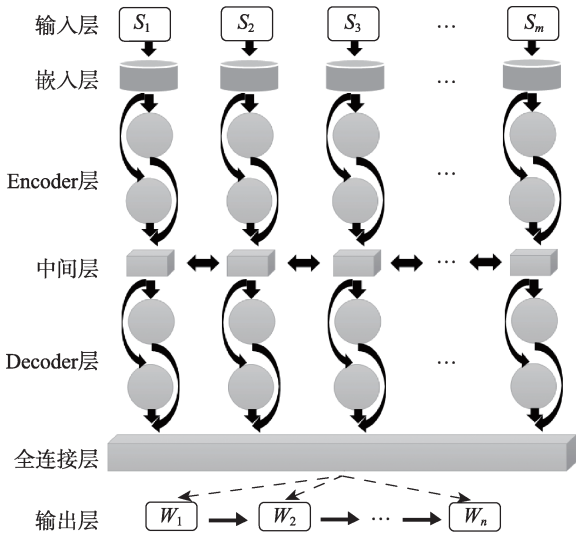

图 3　基于 Encoder-Decoder 的标题生成模型

以上为使用 seq2seq Encoder-Decoder 模型实现标题生成的概要流程，这一架构的序列语言模型在诸多其他任务上也都取得了较好的效果。但其也存在一定弊端：①Encoder 将输入编码为固定大小状态向量的过程实际上是一个“信息有损压缩”的过程，转化向量过程中信息的损失率和信息量的大小呈正相关。②随着 sequence length 的增加，较长时间维度的序列输入会引起 RNN（recurrent neural network）模型的拟合中出现梯度弥散。针对上述问题，本文采用信息密度更为富集的摘要代替全文作为输入，并引入 Attention 机制加以辅助解决。尽管如此，模型效果仍有巨大的提升空间，后续研究中将进一步引入关键词特征信息进行问题与方法的识别。

3.3　生成标题的后续处理

在实现特定样式标题的生成后，本研究需要对所得到的生成结果进行分词、句法分析以及词性标注等后续处理，最终应用基于模板抽取的方法从标题中识别出相应的问题方法指代词，完成学术文本中问题方法词的识别获取。

表 1 给出了所生成标题中频次最高的 5 种组合形式，以及对应的问题方法词抽取规则。其中，对于形式为“基于 A 的 B 的 C”的标题较为特殊，涉及 3 个对象主体。本文依循逻辑推理将 A 与 C 认定为该文的核心研究方法和核心研究问题（A 和 B 是

方法问题对应关系，B 和 C 是方法问题对应关系）。此外，通过表 1 中的统计结果可发现，本文所提出的标题生成模型能够较好的学习标题的样式规则特征，使得所生成标题能够满足本文的任务需求。最后，为避免因不同抽取规则造成的实验效果波动，依据生成标题的统计结果，选用占比最高（97%）的规则模板“基于 A 的 B”统一完成所有生成标题中的问题方法抽取，即视 A 为文中的研究方法，B 为研究问题。

表 1 标题统计及问题方法抽取规则

排序	组合形式	频次	抽取规则
1	基于 NP 的 NP	68961	基于<Method>的<Question>
2	基于 NP VP 的 NP	12818	基于<Method>的<Question>
3	基于 NP 的 NP 的 NP	8495	基于<Method>的 NP 的<Question>
4	基于 NP 的 VP NP	5341	基于<Method>的<Question>
5	基于 VP NP 的 NP	5045	基于<Method>的<Question>

4 实验与结果分析

4.1 实验环境

本研究的所有实验均在表 2 所示的环境配置中完成。

表 2 实验环境

实验环境	环境配置
操作系统	Ubuntu16.04LTS
GPU	NVIDIA GeForce GTX 1080 Ti
内存	16 G
编程语言	Python3.6
深度学习框架	Keras2.2.4
word embedding 训练工具	word2vec[36]

4.2 实验评价及指标

本文是在生成特定样式标题的基础上，应用规则匹配实现学术文本中问题方法指代词的识别获取。因此，本次实验及评价涵盖两方面：标题的生成质量和问题方法的命中效果。由于当前研究多为基于有监督学习的判别式分类，少有采用生成式的策略实现词汇功能的自动识别，故在文中并未设置对照实验。

为了能够对标题的生成质量以及问题方法的命中效果进行全面评估，本文共选取了四项评价指标：BLEU、Turing test、Exact match 和 Unigram。

Exact match 是检索领域中一种常用的关键词匹配模式，要求匹配项之间的字符完全相同；Unigram 是在单个字符层面计算匹配项中出现相同字符的比率；BLEU 是一种基于 *N*-gram 均值的相似度计算方法，被广泛应用于机器翻译评价中[37]；Turing test 则是一种验证机器是否具备人类思维的著名测试，旨在消除机器与人类之间的模糊性，在本文中用以衡量标题生成模型的学习能力[38]。

具体而言，在标题生成质量评价上，使用 BLEU 和 Turing test 在语句级层面评测所生成标题的信息度和流畅度；在问题方法命中评价上，使用 Exact match 和 Unigram 在字符级层面评测问题与方法的命中率。

4.3 实验数据集及参数设置

本文的实验数据来自百度学术和 Google 学术，选取工程技术、计算机和图书情报等多个领域的 2000—2018 年中文学术期刊论文共 574752 篇。经规则过滤后，得到标题样式为“基于 A 的 B”中文期刊文献共计 163367 篇（占比约 28%），其中每篇文献包含文章标题及摘要字段。对数据集乱序处理后从中等比例随机抽取出 4000 篇文献作为测试集，其余则作为训练集用于模型拟合。

训练参数设定上，本实验选择生成式文本摘要任务常用的预设初始值并经多轮迭代调优后：神经网络隐藏层维度为 128；嵌入层向量化维度设为 300（未使用预训练词向量）；词汇表（Vocab）mini_count 为 32；字符最大长度为 400；训练最小批量为 64；迭代 epoch 次数为 100，学习率采取衰减策略（初始值为 1e-5，每训练 500 步衰减 5%）。

4.4 实验结果及分析

4.4.1 标题生成质量评测

语句层面的标题生成评测需要同时考虑词位信息和语义信息，如标题的可读性和信息还原度。因此，本文选择 BLEU 和 Turing test 两种指标对标题的生成质量予以量化评价。

1）BLEU

BLEU 的思想是判断源标题与生成标题的相似度，其原理是计算两个标题中 *N* 元共现词的频率，并依据 *N* 值（*N*=1, 2, 3）进行加权求和。一般而言，1-gram 用以表示对原文信息的还原度，2-gram 和 3-gram 则反映语句的流畅性和可读性。BLEU 具体计算公式为

$$\mathrm{BLUE} = \mathrm{BP}\cdot\exp\left(\sum_{n=1}^{N} w_n \lg P_n\right)$$

其中，BP（brevity penalty）为引入的惩罚因子，用于修正*N*-gram匹配值与句子长度间的负向关系；P_n为*N*-gram下的计算得分，w_n为其对应权重值，通常为1/*n*。在本次BLEU测评中，选用测试集中的全部数据共计4000条，用以与标题生成模型的结果进行BLUE匹配计算。表3为BLEU测试的详细结果。

表3　BLEU测试结果

1-gram	2-gram	3-gram	BLEU
0.640	0.501	0.390	0.424

由表3分析发现，1-gram、2-gram和3-gram的结果呈依次单调递减状态：1-gram最高，为0.640；3-gram最低，为0.390。然而，由于原标题与生成标题依循相同的样式特征——即均含有“基”“于”和“的”这三个特定字符，因此，1-gram结果的单独参考意义相对有限。通过差值比较分析发现，即使在1-gram结果略显“虚高”的情况下，标题在2-gram上的测试表现较1-gram并未出现较大程度的下滑，1-gram、2-gram及3-gram的测试成绩以相对平滑的幅度层级递减。该实验结果表明，由本文所设计的标题生成模型能够在信息完整度与语句流畅性上较好的满足需求。

2）Turing test

Turing test最初被用于判定机器能否表现出与人等价或无法区分的智能，在本文中用于衡量标题生成模型在模拟人类写作上的学习能力。具体而言，本文采用文献[38]中的Turing test测试方法：为每一段摘要配对两个标题——原文标题和机器生成标题，在未告知的情况下由三名博士研究生依据摘要内容进行最优标题投票，选择票数≥2的标题作为最终结果。

在表4所示的Turing test实验样例中，标题1和摘要均为原文内容（由人类撰写），标题2则为对应机器生成结果。限于人工评测方法的既有缺陷，本次Turing test实验只随机选取了200条数据作为测试集，具体结果如表5所示。从表5结果可发现，在大多数情况下（65%），模型生成标题的质量在一定程度上能达到原文水准，少部分情况下表现更优（28%）。该实验结果表明，基于Encoder-Decoder架构的seq2seq模型能够较好的学习人类在标题上的行书特征，可为学术文本中核心问题与核心方法的识别研究提供有力支撑。

表4　Turing test样例

标题1	基于均值聚类分析和多层核心集凝聚算法相融合的网络入侵检测
标题2	基于K-Means算法的网络入侵检测
摘要	为了提高网络入侵的检测率，以降低误检率，提出一种基于均值聚类分析和多层核心集凝聚算法相融合的网络入侵检测模型。利用K-Means算法获取多层核心集凝聚算法的核心集，用其替代原粗化过程得到的顶层核心集，实现顶层核心集的快速准确定位，简化算法的计算复杂性。将KM-Mul CA算法应用到入侵检测模型，采用KDD CUP 99数据集进行仿真实验。结果表明，该模型可以获得理想的网络入侵检测率和误检率。

表5　Turing test实验结果

Task	Result	
Turing test	选择原标题（由人撰写）	70（35%）
	选择生成标题（由模型生成）	56（28%）
	标题趋同，难以选择	74（37%）

表6　问题方法命中评测结果

	Unigram	Exact match
问题	0.481	0.254
方法	0.512	0.289

4.4.2　问题方法命中评测

问题方法的命中效果评价需在字符级和词汇级层面，对得到的问题方法词进行真实值匹配计算。因此，本文选择Exact match和Unigram作为评测指标，以代替传统抽取式方法中所选用的准确率、召回率和F1值。问题方法的命中评测结果如表6所示。其中，使用Unigram在单个字符粒度层面测试问题方法词的命中效果；使用Exact match测试模型能够在多大程度上对原标题中的问题方法词予以还原。

从表6发现，更为严格的匹配规则使得Unigram与Exact match的实验结果间存在显著差距。其中，Unigram在问题方法上的结果均值为0.497，Exact match的结果均值为0.272，这表明模型具有以相同字段命中问题和方法的能力。此外，问题和方法在命中效果上的表现也不尽相同：方法的命中均值为0.401，高于问题的命中均值0.368。经分析发现，

其原因是问题和方法在语言层面上的描述差异。通常而言，研究方法相对于研究问题具有更好的表述规范性。例如，对于计算机领域中大多数技术方法，往往能找到既有的约定术语或通用名称，模型在迭代学习后就能够较好的拟合其概率分布。而对于研究问题，开放性的语言组织使得问题的描述形式显得更为多变和复杂，使得其特征学习更为困难。

4.4.3 综合评测

鉴于以上评测方法均存在一定缺陷，本文采用了量化评分的方式对生成标题的质量以及问题方法的命中进行综合评价。Unigram 和 Exact match 无法识别问题和方法的同义词及变体，如 SVM 与支持向量机虽指向同一实体，但 Unigram 与 Exact match 两种指标均无法对其匹配。同时，Turing 测试中无法指定可依循的评测规则，掺杂了较高主观性。因此，本文从五个层面（表 7）对标题的生成质量和问题方法的命中效果进行综合评测。具体流程如下：①从测试集中随机选出 500 条数据，每条数据包含标题和摘要字段；②将 500 条数据中的原标题均替换为对应的机器生成标题，并在未告知的情况下由三名博士研究生进行独立评测；③要求在理解摘要语义的基础上完成每个待测标题的量化评分；④独立重复多次实验，对结果累计求均值。综合评测的最终结果如图 4 所示。

从图 4 中生成标题在得分序列上的分布可知，

表 7 综合评测评分细则

序号	标题评测标准	分值
1	标题通顺流畅，无语法问题	1
2	能够揭示文中的研究问题	1
3	能够揭示文中的研究方法	1
4	能够对文本的核心问题准确描述	1
5	能够对文本的核心方法准确描述	1

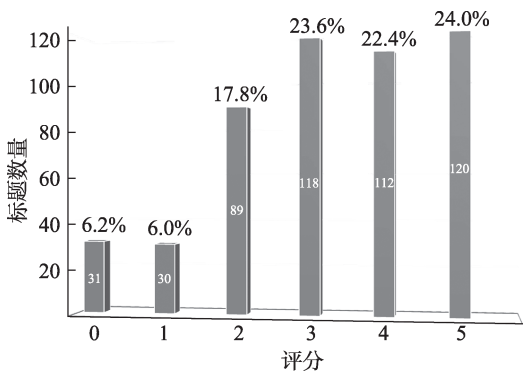

图 4 综合评测结果

生成标题的评测结果集中于 3~5 分区间（70%）。其中，能够准确描述文本问题或方法的高质量标题（分值≥4）占比达到 46.4%。该结果表明通过深度学习方法的应用，本文所提出的基于标题生成策略学术文本问题的方法识别具备相当的可行性和有效性。

由于本文的目的是通过生成特定样式的规则标题实现文本中核心问题与核心方法的获取，与传统标题生成任务[29]或文本摘要任务[31]具有一定区别，因此，本研究并未与之进行对照实验。从表 8 所示的结果样例可发现，对于具有一定行文范式的摘要而言，通过大规模样本的学习，模型能够较好地捕获摘要中的关键语义信息，继而生成限定内容及形式的目标标题。

5 结 语

学术文本词汇功能识别的目的是抽取文本中具有特定意义的表征词汇。受限于数据集等诸多因素的制约，目前基于有监督学习的分类式识别方法存在识别准确率低、召回率有限和泛化性差等问题。因此，本文提出了一种基于深度学习和标题生成策略的文档级学术文本词汇功能识模型，将问题方法指代词的抽取问题转化为特定形式的标题生成问题，在规则标题的基础上实现特定功能词汇的生成和获取。实验结果表明，通过深度学习方法的应用，标题生成策略能够有效识别出描述学术文本研究问题和研究方法的功能性词汇。

本研究仍然存在诸多不足：①学术文本的词汇功能是对词汇在学术文本中角色的定义，包括且不限于问题、方法、领域、工具以及指标等，本文为简化处理，仅仅选取了学术文本中最为核心的问题和方法作为本次的研究对象，后续将采用其他特征和策略实现更为广义的词汇功能识别；②本文仅使用了 LSTM、GRU 等模型，未将 BERT、Transformer 等模型应用于文本信息的语义表征，这些模型的引入能进一步提升识别的效果；③模型仅仅使用了学术文本中的标题和摘要，在语义建模中未能加入关键词、引文网络、作者行文偏好等信息，这些信息的引入对提升模型的效果是有潜在价值的。后续研究将在更大的数据集上开展，应用 Transformer、强化学习等表现力更强的深度学习方法，通过分析文献的类型（技术研究论文、应用研究论文、综述）、引文网络、作者偏好等信息，实现更加精确和鲁棒的词汇功能识别。

表 8　机器生成标题结果样例

样例 1	原标题	基于词性组合规则改进的中文句子极性判断方法
	机器生成标题	基于半监督学习的中文句子情感判断
	摘要	介绍基于词性组合规则改进的中文句子极性判断方法，**提出一种基于半监督学习的中文句子极性判断框架**。在传统的完全基于情感词典方法的基础上，结合词性组合规则这一重要特征对中文句子进行极性判断。首先，分析中文句子中情感短语、情感词语的词性组合规则。然后，将情感短语、情感词语的词性组合规则用于中文句子极性判断。根据词性组合规则集抽取评测句子中的候选情感短语、情感词语；而后，计算句子的情感信息总量和句子的情感值，根据句子的情感信息总量将句子分为主观句、客观句，根据句子的情感值将主观句子分为积极情感句、消极情感句、中立情感句。实验结果证明，该方法在主客观分类上 F 值较高，可以达到 77.4%；在主观句情感分类上，可达到的 F 值为 62.5%。相比较于已有方法，基于词性组合规则改进的中文句子极性判断方法的 F 值有了明显的提高。
样例 2	原标题	基于 DSP 的自适应弱小目标检测方法
	机器生成标题	基于 DSP 的图像目标检测技术研究
	摘要	针对复杂背景下弱小目标检测的难题，**提出一种基于 DSP 的自适应背景预测弱小目标检测新方法**. 该方法在 DSP 为核心的嵌入式图像处理系统平台上，以自适应背景预测算法为基础，在 DSP 集成开发软件 Code Composer Studi0 3.3 上采用 C 语言编写弱小目标检测程序. 根据图像的相邻像素的灰度特性选取不同的背景预测模型对连续四帧原始图像进行自适应背景预测得到背景预测图像，背景预测图像与原始图像相减得到残差图像；对残差图像采用交叉差分算法和自适应阈值分割处理得到二值图像；对二值图像采用逻辑与运算和形态学开运算，获得真实弱小目标. 实验结果表明，该方法可以有效地检测到弱小目标，且与中值滤波算法相比，该算法预处理时间减少 22%，虚警概率降低 6%，检测到的目标面积增大 2.3 倍。
样例 3	原标题	基于均值聚类分析和多层核心集凝聚算法相融合的网络入侵检测
	机器生成标题	基于 K-Means 算法的网络入侵检测
	摘要	为了提高网络入侵的检测率，以降低误检率，**提出一种基于均值聚类分析和多层核心集凝聚算法相融合的网络入侵检测模型**。利用 K-Means 算法获取多层核心集凝聚算法的核心集，用其替代原粗化过程得到的顶层核心集，实现顶层核心集的快速准确定位，简化算法的计算复杂性。将 KM-Mul CA 算法应用到入侵检测模型，采用 KDD CUP 99 数据集进行仿真实验。结果表明，该模型可以获得理想的网络入侵检测率和误检率。

参　考　文　献

[1] Hensiak K. Too much of a good thing[J]. Legal Reference Services Quarterly, 2003, 22(2-3): 85-98.

[2] 孟慧岚, 高鲁山. 科技期刊论文分类标引的探讨[J]. 编辑学报, 2002, 14(1): 27-28.

[3] Ribaupierre H D, Falquet G. Extracting discourse elements and annotating scientific documents using the SciAnnotDoc model: A use case in gender documents[J]. International Journal on Digital Libraries, 2018, 19(2-3): 271-286.

[4] Bikel D M, Miller S, Schwartz R, et al. Nymble: a high-performance learning name-finder[C]// Proceedings of the Fifth Conference on Applied Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 1997: 194-201.

[5] 赵军, 刘康, 周光有, 等. 开放式文本信息抽取[J]. 中文信息学报, 2011, 25(6): 98-110.

[6] 刘怀军, 车万翔, 刘挺. 中文语义角色标注的特征工程[J]. 中文信息学报, 2007, 21(1): 79-84.

[7] 石进, 韩进, 赵小柯, 等. 基于语境概念核心词提取算法研究[J]. 情报学报, 2019, 38(11): 1177-1186.

[8] Abney S P. Parsing by chunks[M]// Berwick R C, Abney S P, Tenny C. (eds) Principle-Based Parsing. Dordrecht: Springer, 1991: 257-278.

[9] Palmer M, Gildea D, Xue N W. Semantic role labeling[J]. Synthesis Lectures on Human Language Technologies, 2010, 3(1): 1-103.

[10] 文勖, 张宇, 刘挺, 等. 基于句法结构分析的中文问题分类[J]. 中文信息学报, 2006, 20(2): 33-39.

[11] Kondo T, Nanba H, Takezawa T, et al. Technical trend analysis by analyzing research papers' titles[C]// Proceedings of the Language and Technology Conference. Heidelberg: Springer, 2011: 512-521.

[12] Nanba H, Kondo T, Takezawa T. Automatic creation of a technical trend map from research papers and patents[C]// Proceedings of the 3rd International Workshop on Patent Information Retrieval. New York: ACM Press, 2010: 11-16.

[13] Trappey A J C, Trappey C V, Govindarajan U H, et al. A review of technology standards and patent portfolios for enabling cyber-physical systems in advanced manufacturing[J]. IEEE Access, 2016, 4: 7356-7382.

[14] Choi S, Yoon J, Kim K, et al. SAO network analysis of patents for technology trends identification: a case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells[J]. Scientometrics, 2011, 88(3): 863-883.

[15] Cheng T Y, Wang M T. The patent-classification technology/function matrix-A systematic method for design around[J]. Journal of Intellectual Property Rights, 2013, 18(2): 158-167.

[16] Gupta S, Manning C D. Analyzing the dynamics of research by

extracting key aspects of scientific papers[C]// Proceedings of 5th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2011: 1-9.

[17] Tsai C T, Kundu G, Roth D. Concept-based analysis of scientific literature[C]// Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. New York: ACM Press, 2013: 1733-1738.

[18] 程齐凯. 学术文本的词汇功能识别[D]. 武汉: 武汉大学, 2015.

[19] 李信, 程齐凯, 刘兴帮. 基于词汇功能识别的科研文献分析系统设计与实现[J]. 图书情报工作, 2017, 61(1): 109-116.

[20] 刘智锋, 李信, 程齐凯, 等. 学术文本关键词语义功能数据集构建与分析——以 Journal of Informetrics 为例[J/OL]. 图书馆论坛, 2019, 39(7): 64-74.

[21] Jin R, Hauptmann A G. Automatic title generation for spoken broadcast news[C]// Proceedings of the First International Conference on Human Language Technology Research. Stroudsburg: Association for Computational Linguistics, 2001: 1-3.

[22] 李浥尘, 胡珀, 王丽君. 基于神经网络的体育新闻自动生成研究[J]. 中文信息学报, 2018, 32(3): 77-83.

[23] 李勇, 成红红, 梁新彦, 等. CNN 图像标题生成[J]. 西安电子科技大学学报, 2019, 46(2): 152-157.

[24] Zeng K H, Chen T H, Niebles J C, et al. Title generation for user generated videos[C]// Proceedings of the European Conference on Computer Vision. Cham: Springer, 2016: 609-625.

[25] 汤鹏杰, 谭云兰, 李金忠, 等. 密集帧率采样的视频标题生成[J]. 计算机科学与探索, 2018, 12(6): 981-993.

[26] Ribeiro R, Matos D M D. Extractive summarization of broadcast news: comparing strategies for European Portuguese[C]// Proceedings of the International Conference on Text, Speech and Dialogue. Heidelberg: Springer, 2007: 115-122.

[27] Nallapati R, Zhou B W, dos Santos C, et al. Abstractive text summarization using sequence-to-sequence RNNs and beyond[C]. Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. Stroudsburg: Association for Computational Linguistics, 2016: 280-290.

[28] Nallapati R, Zhai F F, Zhou B W. SummaRuNNer: a recurrent neural network based sequence model for extractive summarization of documents[C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2017: 3075-3081.

[29] Ayana, Shen S Q, Zhao Y, et al. Neural headline generation with sentence-wise optimization[OL]. (2016-10-09). https://arxiv.org/pdf/1604.01904.pdf.

[30] Rush A M, Chopra S, Weston J. A neural attention model for abstractive sentence summarization[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015: 379-389.

[31] Chopra S, Auli M, Rush A M. Abstractive sentence summarization with attentive recurrent neural networks[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2016: 93-98.

[32] Scott M, Thompson G. Patterns of text: in honour of Michael Hoey[M]. Amsterdam: John Benjamins Publishing Company, 2001.

[33] Paiva C E, da Silveira Nogueira Lima J P, Paiva B S R. Articles with short titles describing the results are cited more often[J]. Clinics, 2012, 67(5): 509-513.

[34] Jamali H R, Nikzad M. Article title type and its relation with the number of downloads and citations[J]. Scientometrics, 2011, 88 (2): 653-661.

[35] Putra J W G, Khodra M L. Automatic title generation in scientific articles for authorship assistance: A summarization approach[J]. Journal of ICT Research and Applications, 2017, 11(3): 253.

[36] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[OL]. (2013-09-07). https://arxiv.org/pdf/1301.3781.pdf.

[37] 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. Stroudsburg: Association for Computational Linguistics, 2002: 311-318.

[38] Wang Q Y, Huang L F, Jiang Z Y, et al. PaperRobot: incremental draft generation of scientific ideas[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 1980-1991.

（责任编辑 魏瑞斌）