情报学报　2020 年 12 月　第 39 卷　第 12 期
Journal of the China Society for Scientific and Technical Information, Dec. 2020, 39(12): 1320-1329

DOI: 10.3772/j.issn.1000-0135.2020.12.008

学术文本词汇功能识别——基于BERT向量化表示的关键词自动分类研究

陆　伟[1,2]，李鹏程[1,2]，张国标[1,2]，程齐凯[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　关键词作为学术文本中映射全文主题内容的词汇或术语，能够为知识精准检索和文本大规模计算提供重要的底层语义标签。当前学术文本中的关键词存在使用意图不明、语义功能模糊及上下文信息缺失等问题。为此，本文提出了一种基于有监督学习的神经网络方法，对关键词所承载的语义功能进行分类，实现对学术文本中研究问题和研究方法的识别。本文以计算机等领域为期10年的学术期刊论文为训练语料，利用BERT及LSTM方法构建分类模型，实验结果显示，本文所提出的方法较传统更优，其整体准确率、召回率和F1值分别达到0.83、0.87和0.85。

关键词　学术文本；关键词；语义功能识别；深度学习

Recognition of Lexical Functions in Academic Texts: Automatic Classification of Keywords Based on BERT Vectorization

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

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

Abstract: As vocabulary or terminology that maps the full-text subject matter content in academic texts, keywords can provide important underlying semantic labels for knowledge retrieval and large-scale text computation. At present, there are problems in the use of keywords in academic texts, such as unclear intention, fuzzy semantic function, and lack of context information. Therefore, a neural network method based on supervised learning is proposed to classify the semantic functions carried by keywords to facilitate the identification of research questions and research methods in academic texts. In this study, journal papers published during a period of 10 years in the field of computer science were used as the training corpus, and the classification model was constructed using BERT and LSTM models. The results show that the proposed method is better than the traditional method. Its overall accuracy, recall rate, and F1 value reached 0.83, 0.87, and 0.85.

Key words: academic text; keywords; lexical function recognition; deep learning

1　引　言

随着科研社区规模的快速扩张和学术文献数量的急剧增长，从海量的学术论文中快速、精准的获取知识越发困难。为应对日益突出的信息过载问题，不同形式的检索工具和检索策略被逐一提出。

收稿日期：2020-05-16；**修回日期**：2020-10-11

基金项目：国家社科基金重大项目“基于认知计算的学术论文评价理论与方法研究”（17ZDA292）。

作者简介：陆伟，男，1974年生，博士，教授，研究方向为信息检索、知识管理、数据智能等，E-mail：weilu@whu.edu.cn；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；程齐凯，男，1989年生，博士，副教授，研究方向为自然语言处理、信息检索、机器学习。

然而，传统的学术信息检索和知识管理多停留于篇章层面的文档集建模，对学术文本深层语义理解的缺失使其无法支持更为细粒度的知识获取服务。Ribaupierre等[1]指出，科研人员的信息获取多为目标和任务驱动，因此更倾向于关注文献中的问题、方法或结果等特定语篇内容。针对于此，学者们试图通过构建词汇粒度的语义功能标签[2]，在理解文本内容语义的基础上为指向性的快速知识索引提供底层支持。

关键词作为一种能够揭示文本主题及核心内容的词汇或术语，对其进行语义功能标识具有重要的理论意义和应用价值。在不同的学术文本中，同一关键词可能蕴含不同的语义角色。如对于文献[3]和文献[4]：在前者中，“支持向量机”表征该文献的主要研究问题；在后者中，“支持向量机”则是该文献用于解决“入侵检测”问题的研究方法。着重精简的关键词存在语义内涵不明、使用意图模糊和上下文信息缺失等问题，易造成关键词查询条件下的检索结果质量难以保证且需再次过滤筛选。为使得关键词更好服务于文献检索和计量分析，本文将以学术文本中的关键词为研究对象，进行词汇功能识别研究。

学术文本词汇功能识别是在词汇粒度层面对学术文本中特定目标信息进行识别、抽取。常见的词汇功能包括问题、方法、数据、指标、工具等，这些功能的显现依赖于学术文本的上下文，同一实体在不同的上下文可能会表现出不同功能。在内涵上与词汇功能相似的概念包括知识元[5]、知识实体[6]以及语义角色[7]等。上述概念同本文所用的学术文本词汇功能具有一定的相似之处：均表现为赋予词汇某一特定含义，并将使用符号完成信息的固化、表示。相较于词汇功能，知识元具有更大的概念外延，体系复杂且相对抽象，对学术文本中概念的角色功能聚焦不足。实体需要显式对应于现实事物且能相互区分，无法用于描述学术文本中的抽象概念和严重依赖上下文的词汇语义信息。

目前，学术文本词汇语义功能的相关研究处于方兴未艾的阶段，且鲜有以关键词为对象来针对性探讨其在文本中所承载的功能角色。鉴于传统机器学习方法中存在的特征构造依赖性强等问题，本文设计并实现了一种基于BERT（bidirectional encoder representation from transformers）向量化表示和LSTM（long short-term memory）网络的学术文本词汇功能识别模型，在理解语义的基础上实现关键词词汇功能判定。在关键词具体的功能类别上，考虑到词汇功能相关研究仍处于探索阶段且功能的具体划分受制于领域等诸多因素，故本文着力于识别关键词的领域无关功能，即将关键词在文献中所承载的角色功能暂定为：研究问题、研究方法和其他。此外，为了克服以往词汇功能识别方法中存在的泛化性差、识别准确率低、识别召回率有限等问题，本文构建了一个包含12万条数据的标注数据集，用于进行模型训练和效果验证。

本文后续内容安排如下：第2节介绍词汇功能识别相关研究的进展现状，第3节详细描述识别方法与深度模型构建，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

实体抽取是信息抽取的重要子任务之一，早期Hearst[8]、Soderland[9]等采用基于规则、字典及模板的方法进行了尝试性探讨，并取得了丰富研究成果。近年来，随着深度学习、预训练模型等统计学习方法的兴起[10]，选择合适的机器学习方法和文本表征模型成为了实体抽取研究的主流思想[11]。例如，SemEval组织的ScienceIE评测任务[12]中，参与者们从多个粒度层面论证了统计学习方法在实体及关系抽取中的有效性。

当前，针对学术文献中理论、算法、术语等科技实体的识别抽取，学者们已经开展了一些工作。陈锋等[13]采用基于CRF（conditional random field）的机器学习方法对学术期刊中的理论实体进行了识别，同时使用语义外部资源对理论实体进行特征泛化，继而解决数据稀疏问题和提高识别准确率。赵洪等[14]在前者的基础上研究了理论术语的特征构造，提出了一种基于弱监督学习的理论术语抽取模型。化柏林[15]分析了情报学领域中方法类术语的表达形式，通过应用词表与规则相结合的方法实现了学术文献中方法术语的抽取。余丽等[16]提出了一种面向细粒度知识元抽取的深度学习模型，并基于改进Bootstrapping技术自动构建大规模的标注语料库，使得模型在“研究范畴”“研究方法”“实验数据”和“评价指标及取值”这四类知识元的抽取上更具有鲁棒性。

得益于自然语言处理技术在结构化文本上的应用和优异表现，文本的词汇功能研究也越发被更多学者所关注。词汇语义功能自动识别最早的研究成果是2009年Kondo等[17]的论文，该文将科研文献标题中的词汇划分至“领域”“问题”“方法”及“其

他”四个功能标签中，通过遍历特定领域内的所有方法/技术来描绘技术的演化路径及未来的发展趋势。随后，Nanba 等[18]使用支持向量机对专利及科研文献中的词汇进行“技术”及“影响”的识别分类，继而构建各领域内的技术趋势图。Gupta 等[19]通过抽取科技文献中的“贡献”“方法”及“领域”来分析该类研究的动态变化。Tsai 等[20]将词汇功能划分为“技术”和“应用”两类，提出了一种无监督的聚类算法实现科研文本中的概念识别与分类。程齐凯[21]在总结前述工作的基础上对学术文本词汇功能进行了明确定义：词汇在学术文本中承担的语义角色，并将词汇功能划分为“领域相关词汇”和“领域无关词汇”。商宪丽[22]从交叉学科的角度构建了学科-对象-方法主题网络模型，以揭示科学知识在各学科之间的流动机理。Heffernan 等[24]认为科技文献是用于描述问题解决活动及过程的文本，提出一种能够对短语进行科学问题和解决方法二元决策的自动分类器，继而实现文献中问题和方法的获取。随着词汇语义功能研究的不断推进，学者们对词汇功能的实际应用也进了初步尝试。Huang 等[25]所构建的 AKMiner（academic knowledge miner）系统能够自动识别学术文本中“概念”及“关系”功能词汇，并利用可视化的方式对生成的知识图谱进行描绘；李信等[26]从学术文本词汇功能的角度出发，考虑科研文献中词汇的语义功能，设计和实现了一个基于词汇问题与方法识别的科研文献分析系统。

从现有成果看，细粒度知识单元的识别和抽取已经得到了各个领域的关注，学者们从多个视角对科技文献中的词汇功能进行了一些尝试和探讨，并获得了持续性的进展。但总体而言，起步较晚的词汇功能研究仍位于探索阶段，存在诸多不足之处，具体表现在：①现有研究多采用传统机器学习方法，对于人工构造特征的强依赖性使得最终的识别效果难以保证；②缺乏面向词汇语义功能的公开数据集，使得相关研究的开展颇受掣肘。为了解决上述问题，本文提出了一种基于深度学习的词汇功能识别模型，以学术文本中的关键词为研究对象进行词汇语义功能的判别，依据最终的判别结果实现文本中研究问题与研究方法的获取，继而为精准知识检索、语步结构分析和大规模文本计算等下游任务提供底层索引支持。同时，本文自行采集、构建 12 万条标注数据集以完成模型的训练和拟合，从而提高词汇语义功能判别模型的准确性和可靠性。

3 研究方法

3.1 词汇功能框架

词汇功能，即词汇在特定上下文环境中所承载的语义功能角色[21]。目前而言，学者们并未就学术文本词汇功能的类别体系达成共识（表 1）。一般而言，问题和方法在多数词汇功能划分体系中都有所体现，但此外的功能设定差异性较大。出现多样的词汇功能设定，一方面是由于不同学者会因为各自的应用目标给出针对性的词汇功能划分；另一方面，则是由于不同学科会有各自领域相关的词汇功能类别。例如，工具、评测指标和数据集更易频繁出现于计算机学科文献中，理论数学的研究文献则并不适用于“测评指标”“数据”此类的功能定义，而代之以公理、公式等功能类别。

表 1 词汇功能划分表

来源	功能划分	研究目的
Kondo 等[17]	Head；Goal；Method；Other	描述技术路径的演化
Nanba 等[18]	Technology；Effect	构建技术趋势图
Gupta 等[19]	Focus；Technique；Domain	分析特定研究的动态变化
Tsai 等[20]	Focus；Application	概念识别与分类
Huang 等[25]	Concept；Relation	生成概念知识图谱
程齐凯[21]	领域相关词汇：研究问题、研究方法 领域无关词汇：工具、数据、指标等	词汇功能框架搭建
商宪丽[22]	学科；对象；方法	多学科间的知识流动描绘
Heffernan 等[24]	Problems；Solutions	构建“问题-方法”对

鉴于词汇功能与所在领域的密切关系，本文沿用文献[21]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两种类别。其中，领域无关词汇功能仅包含两类：问题和方法。学术文献作为科研工作及成果的反映和固化，一般都有明确的研究问题并尽可能给出解决方案。通过梳理表 1 所示的已有词汇功能划分发现，尽管学者们对于具体功能的划分存在分歧，但在“研究问题”与“研究方法”的功能认同上却表现出一致的肯定。领域相关词汇功能是指具有领域适用性的功能词汇。针对不同的学科领域，应定义相应所配套的功能类别。如在计算机领域中，领域相关词汇功能应包含数据、工具、评价指标等；在医学领域中，领域相关词汇功能则应包含

疾病、药物以及基因等（图1）。

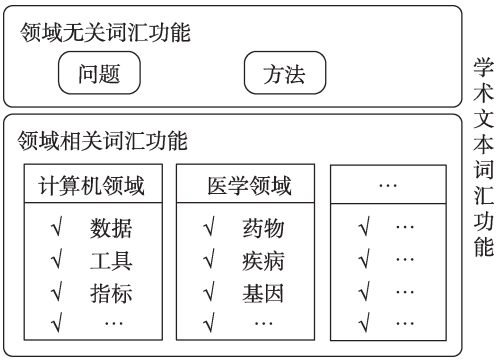

图1　学术文本词汇功能分类体系[21]

学术文本词汇功能识别的本质是信息抽取问题，通过识别承载特定语义功能的词汇来实现文本中研究问题、研究方法及研究对象等目标信息的获取。依据已有的技术方案和任务形式，词汇功能识别可分为两类：基于预设类别的分类标注和基于限定内容的文本生成。

基于预设类别的分类标注。在预先设定词汇功能类别的基础上实行分类或序列标注任务，采用有监督或半监督的训练方式实现所构建分类器的学习拟合，使其能够对学术文本中的词汇进行语义功能的判定。随着机器学习和NLP（natural language processing）等技术的不断发展，分类器在选择上也逐步从依赖于特征构造的SVM（support vector machine）、RF（random forest）等传统机器学习模型过渡至端到端的深度学习模型。

基于限定内容的文本生成。基于文本生成的学术文本词汇功能识别可视为自动文本摘要的下游任务，旨在通过约束序列语言模型输出文本的内容及样式，最终实现目标功能词汇的获取。依据生成策略，自动文摘可分为抽取式和生成式两种，前者是对文档中的词或句进行重要性排序，后者则是在理解文本语义的基础上实现对原文的复述。在过去的数十年里，抽取式的效果通常优于生成式。伴随深度神经网络研究的兴起，基于神经网络的生成式文本摘要得到快速发展，并在众多自然语言处理任务中获得了不俗的表现[23]。

科学研究的本质是一种问题解决（problem-solving）的行为活动[24]，研究问题与研究方法作为大多数科研工作的出发点和着陆点，通常能够更直观地对文献研究内容和研究价值最大化展示。基于以上，本文并未着力于穷尽、列举词汇功能中的所有类目，而是聚焦于学术文本领域无关词汇功能——研究问题与研究方法的识别。通过构建基于有监督学习的词汇功能识别模型，在理解词汇语义信息的基础上挖掘其在文本中所承载的功能性角色，对学术本文中的关键词实现问题方法的判别。

3.2　数据集构造

相较于传统的机器学习方法，深度学习模型对数据集的依赖更为突出和强烈，训练语料的质量与体量直接关联模型的性能效果和泛化能力[27]。目前，尚无公开、大规模的学术文本词汇功能标注数据集，已有的数据集往往数量太少，无法应用于深度学习模型训练。本文通过数据采集、过滤、清洗和自然标注，构建了一个包含121316条标注样本的数据集。

在数据采集方面，使用百度学术、谷歌学术及搜狗学术等网络搜索引擎为数据的获取源，通过手工方式从《计算机学报》《计算机工程》及《情报学报》等多本计算机及图书情报领域期刊中获取2009—2018年刊载的所有文献，从中剔除缺少关键词字段的论文，共获得122544篇文献的相关数据。

为克服传统人工手动数据标注过程中存在的高时耗和高成本问题，本文采用了一种基于规则标题的数据标注方法[17]。在现有的中文期刊中，存在大量标题类似于“基于A的B研究”学术文本。同样，该类特征也显著表现于国外期刊文献中，如ACL、ACM数据库中的收录论文中存有大量“A based on B”“A using B”以及“A for B”的样式标题。考虑到这种规则化的标题在相当程度上是对文中研究问题和方法的揭示，本文将通过该规则特征完成训练语料的数据标注（如图2所示）。具体过程如下：①通过模板匹配从122544篇文本中筛选出40368篇标题为“基于A的B的研究”；②使用Stanford NLP工具包[28]对文本标题进行分词、词性标注和实体抽取；③使用最长字符串匹配法从标题中抽取出问题词与方法词，构造字符串切分树及同义词词典，计算关键词和抽取词的相似度[21]，低于阈值的关键词将被归入类别“其他”。最终，共得到121316条基于规则的标注数据。图3给出了一篇科技文献的标注样本示例，其中输入和输出分别采用“摘要+关键词”和“类别ID”的标注形式。

为了评测模型的真实性能，从获取的原始数据集中随机抽取1000篇文献，由1位情报学博士研究生和2位情报学硕士研究生对该文献及所含的关键词进行两轮标注。第一轮由个人完成，对于个人不

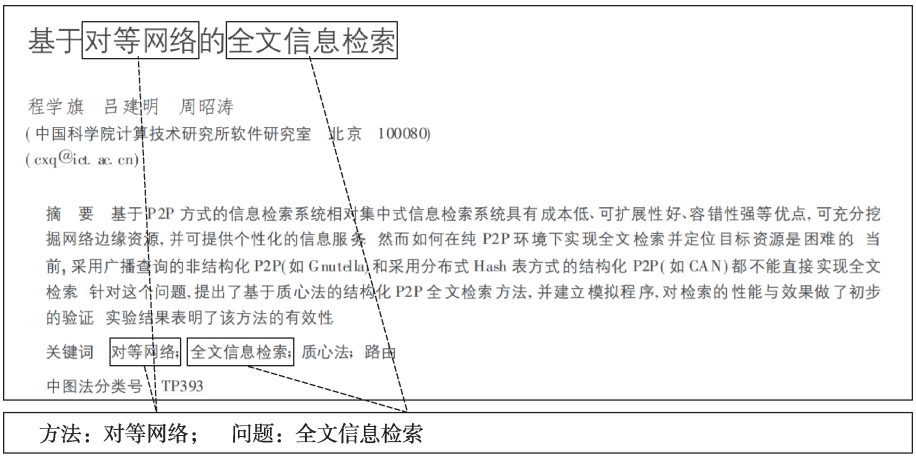

图2 数据标注方法示意图

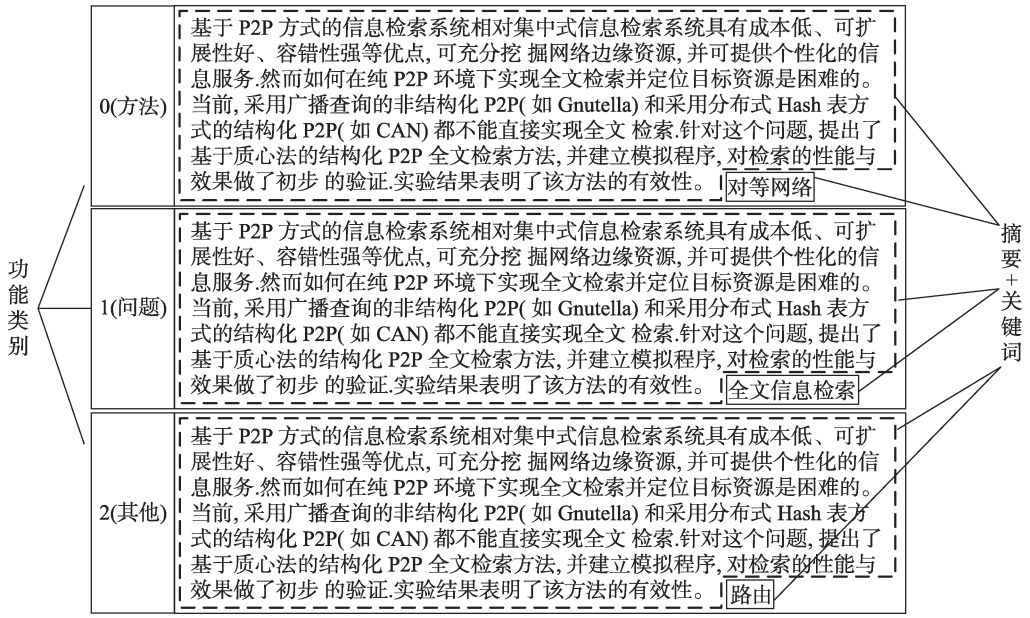

图3 标注样例示意

能确定的数据，在第二轮由3位同学投票决定关键词类型，最终得到4077条标准数据。其中问题类关键词1176个（28.8%），方法类关键词1366个（33.5%），其他类关键词1535个（37.7%）。此外，同时含有问题或方法类关键词的文献为893篇（89.3%）。该结果表明，虽然关键词的列选具有一定主观性，少数文献的关键词可能并未涉及其研究问题或研究方法，但总体而言，问题与方法作为文献的核心研究内容，多可在关键词上得以体现。

3.3 关键词语义功能识别模型

关键词语义功能识别的整体流程如图4所示：①数据集经预处理后将论文的摘要字段与关键词字段输入深度学习模型中；②使用BERT预训练模型

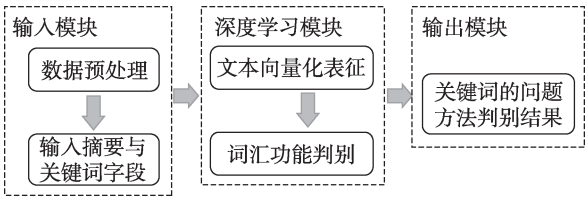

图4 关键词语义功能识别流程图

对输入的文本进行向量化表征，从多个粒度层面捕获摘要与关键词间的关联语义信息，通过LSTM对关键词进行自动分类；③输出关键词的问题方法判别结果。下面对上述过程中所涉及的一些主要方法技术进行简要介绍。

（1）BERT。BERT是由Google AI团队于2018年所提出的一种基于Transformer模型[29]的预训练向

量表示方法。BERT 网络模型在遵循词嵌入一般思想的基础上，进一步增加了词向量模型的泛化能力，通过字符级、词汇级以及句子级的多粒度特征关系挖掘，使得能够对文本的词性、句法和语义等信息进行充分描述。相较于传统词向量模型，BERT 网络模型有以下创新：①引入 position encoding 来描述序列位置信息。对于序列中的每一个元素给定一个待训练的随机初始化词向量，以记录元素在该序列中的位置信息。②选择深层双向的编码层完成词向量的学习。考虑到单词的语义依赖于其所在的上下文环境，即它左右两侧的某些词，采用结合前向和后向的双向 encoding 能够使得词向量具有上下文关联的能力，继而实现动态化的向量词义消歧。③采用 Transformer 作为模型的基本构架。在 encoder 的选择上，BERT 网络模型使用了并行性更优的 Transformer 以取代传统方法中的 RNN（recurrent neural network），多层可叠加的 self-attention 机制使得 BERT 模型能够无视空间和距离学习序列中的词位交互信息。基于上述优点，本文将在嵌入层中应用 BERT 网络对文本进行向量化表示，从多个粒度层面实现文本语义信息的动态表征。由于其强大的文本语义表达能力，BERT 模型在多个自然语言处理任务中得到应用，取得了较好的应用效果。

（2）LSTM。LSTM 是循环神经网络 RNN 的重要实现技术[30]。通过将上一时间步的输出结果重新作为当前时间步的输入特征，RNN 模型能有效地捕获与利用序列数据的历史关联信息，从而实现信息的持久化记忆。虽然理论上 RNN 能够对任意步长的序列数据进行处理，然而由后向反馈机制所引起的梯度消失使得 RNN 在实际中难以处理较长的序列。为了克服这一缺陷，学者们在原有的隐藏层中引入了记忆机制和遗忘机制，以解决传统 RNN 模型中的长期依赖问题[31]。输入门、遗忘门和输出门的共同协作使得 LSTM 能够存取序列的长期依赖信息。当输入门、遗忘门及输出门同时保持开启状态时，遗忘门负责选择性保留和丢失细胞状态中的原有信息，并经由输入门确定细胞状态的信息更新，最终输出门会基于当前的细胞状态生成该时刻的状态输出。

BERT+LSTM 模型能够较好地解决文本表征过程中语义缺失问题，弥补传统 RNN 在长距离上下文信息交互中的不足，可为词汇功能的准确识别提供有力支撑。在具体的模型构造上，本文中的学术文本关键词语义功能识别模型沿用了典型的文本分类结构框架，共分为四层：输入层、嵌入层、中间网络层以及输出层，具体结构如图 5 所示。

在输入层中，选用信息量更为富集的文献摘要代替全文作为关键词的语境上下文，使得能够尽可

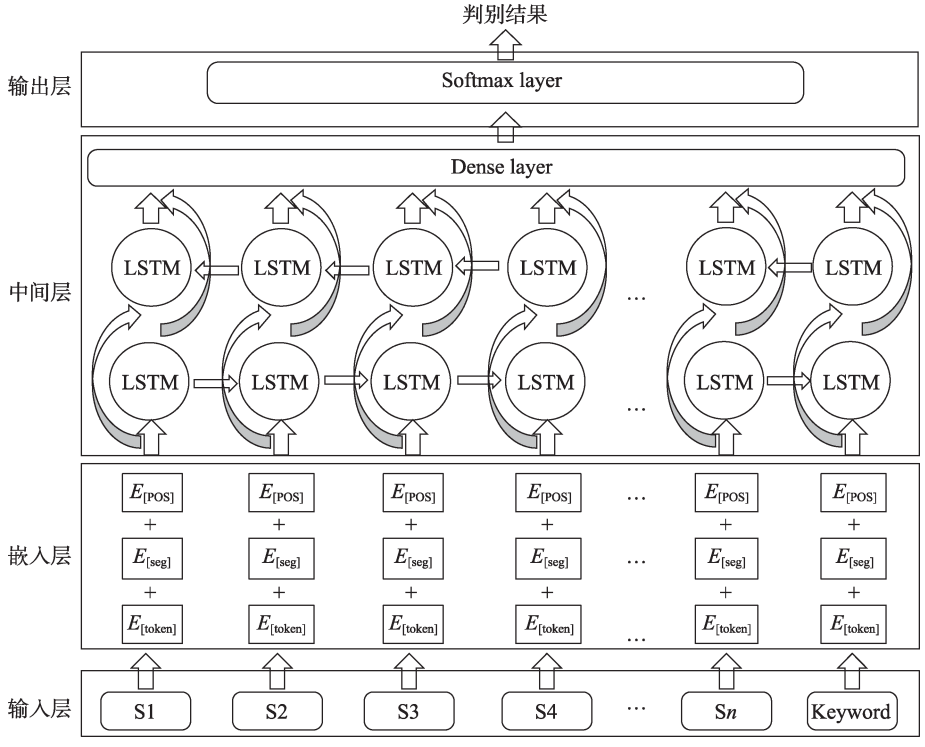

图 5　关键词语义功能识别模型结构图

能保留关键信息的同时显著降低计算时耗。对于每篇待进行词汇功能识别的学术文献，将其摘要 $S_{\text{abstract}}=\{W_1,W_2,\cdots,W_n\}$ 依次与关键词列表 $L_{\text{Keyword}}=\{\text{Keyword}_1,\cdots,\text{Keyword}_m\}$ 中的每个关键词进行拼接，得到关于该文献的模型输入序列集 $S_{\text{input}}=\{S_1,S_2,\cdots,S_m\}$，其中模型的单个输入序列 $S_m=\{W_1,W_2,\cdots,W_n,\text{Keyword}_m\}$ 如图 3 所示。

嵌入层的作用是对文本序列进行向量表征，使得计算机能够读取和理解文本的潜在语义信息。在嵌入层中，使用 BERT 网络完成文本序列向多维向量空间的映射，BERT 自身携有的句子级负采样学习模式能够显著削弱模型对于数据预处理层中分词效果的依赖。经 BERT 网络模型实现输入信息的多粒度动态表征后，文本的最终向量化表示由词条嵌入（token embedding）、分割嵌入（segment embedding）以及位置嵌入（position embedding）三个部分拼接构成。随后，将该状态向量输入至由 LSTM 网络和全连接网络构成的中间层，通过 LSTM 中的前向迭代学习完成文本的上下文信息交互和潜在语义捕获。

最终在输出层，使用 Softmax 分类器对中间层的输出特征向量进行概率分布计算，得到当前关键词的词汇功能标签结果，其中 0、1、2 分别对应功能类别中的方法、问题以及其他。

4 实验与结果分析

4.1 实验环境及评价指标

本次实验的操作系统为 Ubuntu 16.04 LTS，使用 python3.6 和 tensorflow1.9 作为实验的开发环境。实验过程中，选择由 Google 已完成预训练的 BERT 简体中文词向量模型。

在评价指标上，选择查准率（precision）、召回率（recall）及 F1 值来衡量语义功能识别模型的实验效果。其中，F1 值为查准率和召回率的调和平均数，用于评价模型的综合性能，计算公式为

$$\text{F1}=\frac{\text{precision}+\text{recall}}{2}$$

此外，为了能够真实反映出本文所提出模型的判别性能，笔者使用人工手动校对的方式获取了 4077 条标准数据（经多轮验证准确率高于 95%），以作为本模型的实际评测效果参考。

4.2 优化策略及参数设定

为解决过拟合问题，实验将采用正则化和 Dropout 方法以降低模型的复杂度。通过在 LSTM 层前后添加 Dropout 网络层以及在损失函数中引入 L1、L2 正则项，在提升测试集准确率的同时使得模型更加鲁棒。此外，应用 Adam 梯度下降法加快模型的收敛速度。

在训练超参数设定上，选择文本分类任务常用预设初始值并经多轮迭代调优后，BERT 向量化维度设为 768，读取字符最大长度为 512，LSTM 隐藏层神经元数设为 128，训练最小批量为 128，迭代 epoch 次数为 40，学习率指数采取衰减策略（初始值为 2e−5，每训练 500 步衰减 5%），dropout 设为 0.5。

4.3 实验结果及分析

本文分别选用了 SVM、LSTM 和 BERT 三种方法作为对照实验，以检验所设计的 BERT+LSTM 结构模型在关键词语义功能识别任务中的表现效果。基于自建数据集完成模型训练后，采用 10 折交叉验证方法后得到各模型最终的评测结果，具体情形如表 2 和表 3 所示（规则数据集为基于机器标注的数据集，手工数据集为人工手动标注的数据集）。

表 2 规则数据集测试结果

方法	类别	准确率	召回率	F1
SVM	研究方法	0.45	0.61	0.53
	研究问题	0.42	0.55	0.49
	其他	0.64	0.60	0.62
LSTM	研究方法	0.81	0.90	0.85
	研究问题	0.79	0.76	0.77
	其他	0.83	0.74	0.78
Bi-LSTM	研究方法	0.73	0.86	0.80
	研究问题	0.70	0.73	0.72
	其他	0.81	0.75	0.78
BERT	研究方法	0.91	0.88	0.89
	研究问题	0.79	0.82	0.80
	其他	0.83	0.80	0.81
BERT+LSTM	研究方法	0.91	0.87	0.89
	研究问题	0.74	0.84	0.79
	其他	0.84	0.91	0.87

从各个模型的整体识别效果可发现，SVM 在语义功能识别任务中的表现明显劣于其他三种神经网络模型方法（LSTM、BERT 和 BERT+LSTM），这是传统机器学习方法对于特征构造的依赖性所造成的。本次实验中的四种模型均使用了词向量作为文本特征输入，然而传统机器学习方法缺乏对序列化文本的理解能力，无法关联输入特征的上下文信

表3　人工数据集测试结果

方法	类别	准确率	召回率	F1
SVM	研究方法	0.43	0.66	0.52
	研究问题	0.40	0.51	0.45
	其他	0.60	0.48	0.54
LSTM	研究方法	0.68	0.74	0.71
	研究问题	0.69	0.77	0.73
	其他	0.75	0.82	0.79
Bi-LSTM	研究方法	0.71	0.77	0.74
	研究问题	0.70	0.70	0.70
	其他	0.73	0.71	0.72
BERT	研究方法	0.80	0.82	0.81
	研究问题	0.71	0.71	0.71
	其他	0.73	0.80	0.76
BERT+LSTM	研究方法	0.82	0.77	0.79
	研究问题	0.70	0.72	0.71
	其他	0.85	0.83	0.84

息，因而表现糟糕。为验证该观点，笔者将SVM的输入特征由词向量替换为TF-IDF，其准确率和召回率均获得了近10%的提升（由于仍低于上述三种神经网络方法的最低水平，故未列于文中），该实验结果可认为是前述观点的有力佐证。

通过对比LSTM与BERT+LSTM两种模型的实验结果可知，使用BERT向量化表征方法能够较好地提升词汇功能识别模型的分类性能，加入BERT后的LSTM在研究方法和研究问题的识别F1值分别提高了4%和2%。相对而言，BERT模型与BERT+LSTM模型之间的测试结果相差无几。除去在类别“其他”表现上，BERT+LSTM模型相比于BERT模型有些许提升，两种模型在“研究问题”和“研究方法”上的F1结果差值均保持在0.1左右。笔者认为，其造成原因在于BERT模型的自身强大能力使得LSTM能够带来的提升效果显得微乎其微[32]。作为在多项NLP任务中创出最佳纪录的自然语言处理模型，BERT网络具有强大的语法解析和语义理解能力。通过不断获取文本位置、语法和状态等特征信息，自注意力机制使得BERT模型能够在理解文本语义的基础上实现对文本的向量化表示。因此，重复捕获相同文本特征的LSTM无法对模型的性能进行有效提升。此外，对比LSTM与Bi-LSTM（bi-directional LSTM）的实验结果发现，LSTM的性能效果明显较Bi-LSTM更优，笔者经分析认为输入语料的构成方式是造成该现象的主要原因。由于关键词位于实验语料的末尾，只能前向迭代传播的LSTM能够在末端赋予关键词更大的信息权重，Bi-LSTM虽能经前向及后向对文本的潜在语义信息进行双向捕获，但相较LSTM而言丢失了部分关键词信息。

在各个功能类被的识别结果中，所有模型对于“研究方法”类别的识别准确率和召回率均高于“研究问题”类别，这一现象在BERT+LSTM模型中表现最为突出，识别准确率、召回率和F1的差值分别达到0.17、0.03和0.11。通过分析发现，其主要原因为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，对于特定特征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，模型多次学习到该类特征后就能够较好地对其进行识别和判定。而对于研究问题，开放性的语言组织使得其问题的描述形式显得更为多变和复杂，故判别准确率相对较低。人工标注的数据集测试结果如表3所示，四种模型在各项指标上的表现均有不同程度的下降，但整体结果仍符合由表2得出的最终结论。

5　结　语

本文针对现存的关键词标引意图模糊和语义功能不明问题，设计并实现了一种基于深度学习方法的语义功能识别模型，在遵循现有科研范式的基础上对学术文本中的关键词实现研究问题和研究方法的判别，继而为知识精准检索、知识辅助理解及知识规模计算等服务提高底层索引支持。在实验中，本文基于规则匹配构建了12万条训练语料以实现所设计模型的训练与收敛，并将手工标注的4000余条数据作为测试集，以真实反映模型的实际效果。实验结果表明，本文提出方法所取得效果较传统方法具有显著的提升。

尽管本文所提出的语义功能识别模型能够较好地对学术文本中的关键词进行问题与方法判别，但研究仍存在一些不足之处。首先，学术文本通常会涉及多个研究问题和研究方法，该方法及问题究竟为文中的核心研究要点，还是仅仅作为参考背景而提及，本文并未对其予以区分；其次，将摘要代替全文作为模型输入虽能减少计算负荷，但也遗失了大量的上下文特征，如文本的逻辑结构和引文网络等信息；最后，本文仅仅使用了LSTM、BERT等模型进行分类式判别，未引入条件随机场模型，也未考虑使用seq2seq语言模型实现特定功能词汇的生成，这些模型的引入对于词汇功能的识别研究具

有一定的潜在价值。后续研究将在更大的数据集上展开，以增强识别模型的泛化性和鲁棒性。同时，将更多的文本信息（如引文网络）和先验知识（如作者行文偏好）引入词汇功能识别模型，使得能够实现更为细粒度的学术文本词汇功能识别，继而为“哪些〈问题〉可以被哪些〈方法〉解决”等现实场景提供服务支持。

参 考 文 献

[1] 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.

[2] 钱力, 张晓林, 王茜. 科技论文的研究设计指纹自动识别方法构建与实现[J]. 图书情报工作, 2018, 62(2): 135-143.

[3] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1): 32-42.

[4] 饶鲜, 董春曦, 杨绍全. 基于支持向量机的入侵检测系统[J]. 软件学报, 2003, 14(4): 798-803.

[5] 温有奎, 焦玉英. 知识元语义链接模型研究[J]. 图书情报工作, 2010, 54(12): 27-31.

[6] Ding Y, Song M, Han J, et al. Entitymetrics: Measuring the impact of entities[J]. PLoS ONE, 2013, 8(8): e71416.

[7] 刘挺, 车万翔, 李生. 基于最大熵分类器的语义角色标注[J]. 软件学报, 2007, 18(3): 565-573.

[8] Hearst M A. Automatic acquisition of hyponyms from large text corpora[C]// Proceedings of the 14th International Conference on Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 1992, 2: 539-545.

[9] Soderland S. Learning information extraction rules for semi-structured and free text[J]. Machine Language, 1999, 34: 233-272.

[10] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[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: 260-270.

[11] 刘浏, 王东波. 命名实体识别研究综述[J]. 情报学报, 2018, 37(3): 329-340.

[12] Augenstein I, Das M, Riedel S, et al. SemEval 2017 Task 10: ScienceIE - Extracting keyphrases and relations from scientific publications[C]// Proceedings of the 11th International Workshop on Semantic Evaluation. Stroudsburg: Association for Computational Linguistics, 2017: 546-555.

[13] 陈锋, 翟羽佳, 王芳. 基于条件随机场的学术期刊中理论的自动识别方法[J]. 图书情报工作, 2016, 60(2): 122-128.

[14] 赵洪, 王芳. 理论术语抽取的深度学习模型及自训练算法研究[J]. 情报学报, 2018, 37(9): 923-938.

[15] 化柏林. 针对中文学术文献的情报方法术语抽取[J]. 现代图书情报技术, 2013(6): 68-75.

[16] 余丽, 钱力, 付常雷, 等. 基于深度学习的文本中细粒度知识元抽取方法研究[J]. 数据分析与知识发现, 2019, 3(1): 38-45.

[17] Kondo T, Nanba H, Takezawa T, et al. Technical trend analysis by analyzing research papers' titles[C]// Proceedings of the Language and Technology Conference on Human Language Technology, Challenges for Computer Science and Linguistics. Heidelberg: Springer, 2009: 512-521.

[18] 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.

[19] Gupta S, Manning C D. Analyzing the dynamics of research by extracting key aspects of scientific papers[C]// Proceedings of the 5th International Joint Conference on Natural Language Processing. Asian Federation of Natural Language Processing, 2011: 1-9.

[20] 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. NewYork: ACM Press, 2013: 1733-1738.

[21] 程齐凯. 学术文献词汇功能识别[D]. 武汉: 武汉大学, 2015.

[22] 商宪丽. 基于多模主题网络的交叉学科知识组合模式研究——以数字图书馆为例[J]. 情报科学, 2018, 36(3): 130-137, 150.

[23] 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.

[24] Heffernan K, Teufel S. Identifying problems and solutions in scientific text[J]. Scientometrics, 2018, 116(2): 1367-1382.

[25] Huang S S, Wan X J. AKMiner: Domain-specific knowledge graph mining from academic literatures[C]// Proceedings of the International Conference on Web Information Systems Engineering. Heidelberg: Springer, 2013: 241-255.

[26] 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.

[27] Sun C, Shrivastava A, Singh S, et al. Revisiting unreasonable effectiveness of data in deep learning era[C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. IEEE, 2017: 843-852.

[28] Manning C D, Surdeanu M, Bauer J, et al. The Stanford CoreNLP natural language processing toolkit[C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Stroudsburg: Association for Computational Linguistics, 2014: 55-60.

[29] 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. Red Hook: Curran Associates, 2017: 6000-6010.

[30] Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modeling[C]// Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association, 2012.

[31] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[32] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2018: 4171-4186.

（责任编辑　魏瑞斌）