北京大学学报(自然科学版) 第 59 卷 第 1 期 2023 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 59, No. 1 (Jan. 2023)
doi: 10.13209/j.0479-8023.2022.070

依据《中国图书馆分类法》的英文图书分类探索

蒋彦廷[1,2]

1. 四川省水文水资源勘测中心, 成都 610036; 2. 中共金堂县委党校, 成都 610400; E-mail: jiangyanting@mail.bnu.edu.cn

摘要 针对带有中图分类号的英文图书数据量小以及类别不平衡的问题, 将图情领域的文本增强策略(《美国国会图书馆分类法》到《中国图书馆分类法》的类目映射方法和基于中–英文平行的《汉语主题词表》的语义增强方法)与一般领域文本增强策略(向原始英文文本插入标点或连词)相结合, 旨在增强模型泛化能力。实验表明, 综合后的策略能有效地提高模型在测试集的表现, 正确率和宏 F1 值分别上升 3.61 和 3.35 个百分点, 效果优于其他单一的文本增强方法。最后, 通过 BERT 词向量可视化与词语信息熵计算, 分析出丰富的邻近词和语法上的连缀功能是插入标点或连词方法有效的原因。

关键词 预训练语言模型; 中国图书馆分类法; 类目映射; 汉语主题词表; 文本增强

English Books Automatic Classification According to CLC

JIANG Yanting[1,2]

1. Sichuan Hydrological and Water Resources Survey Center, Chengdu 610036; 2. CPC Party School of Jintang County, Chengdu 610400; E-mail: jiangyanting@mail.bnu.edu.cn

Abstract Faced with lacking of English books annotated with CLC (Chinese Library Classification) label and imbalance data, this paper combines augmentation strategies from library, information and general fields: 1) classification mapping from Library of Congress Classification (LCC) to CLC; 2) semantic enhancement based on Chinese-English parallel thesaurus; 3) punctuation or 4) conjunction inserting to initial texts. Experiments show that combining 4 strategies can optimize the performance of models on test set. Accuracy and Macro-F1 respectively increase by 3.61 and 3.35 percentage points. Comprehensive methods is superior to other text enhancement strategies. By BERT word embeddings visualization and words information entropy computing, this paper inferred that the reason why punctuation or conjunction inserting works was the various adjacent words and connection function in grammar.

Key words pre-trained language models; Chinese Library Classification; classification mapping; Chinese thesaurus; text augmentation.

书籍是承载人类知识思想的重要载体。近年来, 中国进口、加工外文图书的规模相当可观。在纸质图书方面, 截至 2022 年 7 月, 中国图书进出口(集团)有限公司累计采选海外图书超过 184 万种, 月均新增超万种[1]。北京大学图书馆 2022 年上半年加工编目的外文新书约 9800 册[2]。

外文图书的进口给国内图书馆或文献数据库的加工编目带来挑战[3]。与中文图书相比, 外文图书分类编目难度更大。第一个原因, 外文图书分类编目对工作人员的外语水平和对具体领域的熟悉度都有较高的要求。第二个原因, 国内外图书分类体系有差异: 国内大部分书店、图书馆、电子书网站参考《中国图书馆分类法》(简称《中图法》)给图书分类。大部分中文图书在版权页已初步标注《中图法》分类号(简称中图分类号), 大大减轻了图书分类编目的负担。然而许多英语国家出版的图书并未采用《中图法》进行分类。

基于上述背景, 本文利用预训练语言模型 BERT (bidi-rectional encoder representations from transformers), 结合图书情报(图情)领域与一般领域

收稿日期: 2022–05–13; 修回日期: 2022–08–03

的文本增强方法，对依据《中图法》的英文图书分类工作进行探索，以期方便读者索书查阅，提高外文图书的利用率和使用效益，优化图书编目与知识管理。

1 相关工作

1.1 国内外英文图书分类情况

国内外代表性图书馆和文献数据库网站的英文图书分类情况如表 1 所示。《中图法》是新中国编制出版的图书资料分类体系，至 2012 年已经出版第五版简本[4]，包括 22 个一级类目①，250 多个二级类乃至更多的细目。《美国国会图书馆分类法》(Library of Congress Classification，简称《国会图书分类法》)是美国国会图书馆设计的资料分类法，将知识分为 21 个基本大类②。《中国科学院图书馆图书分类法》简称《科图法》，1958 年出版第 1 版，采用阿拉伯数字为类目的标记符号，包括 25 个大类和更多的小类。《杜威十进制分类法》(Dewey Decimal Classification，简称《杜威分类法》)③由美国图书馆专家麦尔威·杜威发明，以 3 位数字作为分类码的开头，将知识分为 10 个大类，至 2004 年已出版至第 22 版。

调查发现，首先，在图书管理实务中，中国内地的大多数图书馆与文献数据库网站都依据《中图法》给英文图书编目。一些机构虽然兼用多种分类法，但在给英文图书编制索书号时，仍主要参考《中图法》，在数据库机读目录(Machine-Readable Catalogue, MARC)中将其他分类号作为次要字段。第二，《国会图书分类法》除在美国广泛使用外，在新加坡、中国的香港和台湾的大学图书馆中也有所应用。第三，英国和中国香港的部分图书馆采用《杜威分类法》。

中国内地主要采用《中图法》给外文图书分类原因之一是《中图法》类目详尽，基本涵盖知识的各领域，并与时俱进。《中图法》还设置“互见分类号”，例如隶属“C 社科总论”的“C8 统计学”与“O1 数学”下辖的“O212 数理统计”。双语对照的读物按前一种语言归类，按后一种语言做互见分类[4]。作为树形分类结构，互见分类能较好地表示跨学科、交叉学科知识。此外，《中图法》还有 L，M，W 和 Y 四个一级类目的字母没有使用，为未来新兴学科领域留有空间[5]。另一个原因是中外文图书采用统一的分类号，能提升检索效率，为科技查新、追踪考察国外科学进展夯实基础。最后，实体书店与图书馆通常在图书分类号的基础上编制索书号。依据《中图法》编制索书号，能方便工作人员上架图书，也方便读者查找图书，减轻熟悉两套图书分类法的记忆负担。

1.2 主题词表相关研究

主题词表又称叙词表，是一种阐释某学科领域相关术语的语义词典，是实现信息智能检索的重要资源[6]。国内规模较大的主题词表有两部：《中国分类主题词表》[7]与《汉语主题词表》。后者 1980 年问世，2009 年由中国科学技术信息研究所重编，包括工程技术、自然科学、生命科学、社会科学四部分。截至 2022 年 7 月初，《汉语主题词表》在线服务系统发布术语词条 131400 个[8]。大部分词条由

表 1 国内外代表性图书馆、文献数据库网站采用的英文图书分类体系

Table 1 English book classification taxonomy of libraries and literature databases

图书馆和文献数据库网站	英文图书依据的分类法
北京大学图书馆、北京师范大学图书馆、吉林大学图书馆、首都图书馆等	《中图法》
中山大学图书馆、上海交通大学图书馆等	《中图法》、《国会图书分类法》
浙江图书馆、中国科学技术大学图书馆等	《中图法》、《科图法》④
中国科学院文献情报中心等	《中图法》、《国会图书分类法》、《科图法》
国家图书馆、武汉大学图书馆、南京图书馆、CALIS 联合目录公共检索系统等	《中图法》、《国会图书分类法》、《杜威分类法》
美国国会图书馆、麻省理工学院图书馆、斯坦福大学图书馆、南洋理工大学图书馆、香港中文大学图书馆等、中国台湾大学图书馆等	《国会图书分类法》
英国国家图书馆、伦敦大学学院、香港大学图书馆等	《杜威分类法》

① http://www.ztflh.com

② https://www.loc.gov/catdir/cpso/lcco/

③ https://www.britannica.com/science/Dewey-Decimal-Classification

④ 据 2022 年 7 月初的调研结果，浙江图书馆图书页面显示分类法为《科图法》，但实际标注的分类号依据的是《中图法》。

号组成，部分术语还涉及多个中图分类号。国际上，著名的主题标题表有美国的《国会图书馆主题词表》(Library of Congress Subject Headings, LCSH)[9]和《医学主题词表》(Medical Subject Headings, MeSH) [10]等。

1.3 基于机器学习的文献分类技术

包括图书、论文以及专利文档在内的文献分类是文本分类技术中的特殊领域。在算法模型方面，支持向量机(SVM)[11]、胶囊神经网络[12]、决策树(DT)[13]、长短期记忆(LSTM)[14]、BERT以及预训练模型及其改进版[15−16]已应用到图书或论文的分类任务中。在分类标签方面，可以分为单标签与多标签分类[15]。在文献语种与分类号方面，目前按照《中图法》对中文文献分类的研究较丰富，相关在线服务平台[8,17]也得以建设，也有依据《国会图书分类法》[18]、《杜威十进制分类法》[13]和Web of Science网站学科分类体系[12]对英文文献分类的探索。目前，涉及跨分类法、跨文献语种问题的探索还较少。

1.4 文本数据增强技术

在数据稀疏的情况下，采取文本增强(data augmentation for text)技术有助于生成训练文本的近似样本，避免过拟合，提高文本分类的效果。文本增强包括回译、随机删词、词序打乱、基于静态或动态词向量的词汇替换[19−20]、适量噪声注入[21]、同类文本交叉重组[22]、引入词汇释义[23]、强化学习[24]以及文本复述[25]等方法。依据特定分类法的图书分类是较为特殊的领域，该领域的文本增强方法还有待探索。

2 英文文献分类与文本增强策略框架

根据中英文图书论文的分类经验[14,18]，当每个文本的输入字段为书名和若干反映主题的关键词时，分类效果基本上达到最佳水平。由于文本较长，图书简介字段中非关键信息较多，对分类的贡献不明显，也不利于模型训练收敛。因此，我们使用基于图的TextRank关键词提取方法[26]，首先从图书简介文本中提取权重靠前的若干关键词，与书名一起作为训练数据。

在分类方法方面，本文基于BERT预训练模型①，结合全连接神经网络(FCN)分类器，实现中图法一级分类号B到X的20类文献分类。将支持向量机(SVM)模型、随机森林(random forests, RF)模型、Fasttext模型[27]、基于114万篇英文文献预训练的SCI-BERT模型②和蒸馏轻量化的DistilBERT模型③作为基线模型。由于BERT等预训练模型会采用Wordpiece算法[28]，将英文单词切分为子词(subword)，因此我们只在文本输入非预训练模型前，使用NLTK工具库④将单词词干化。

本文提出的的英文文献分类与文本增强策略框架如图1所示。

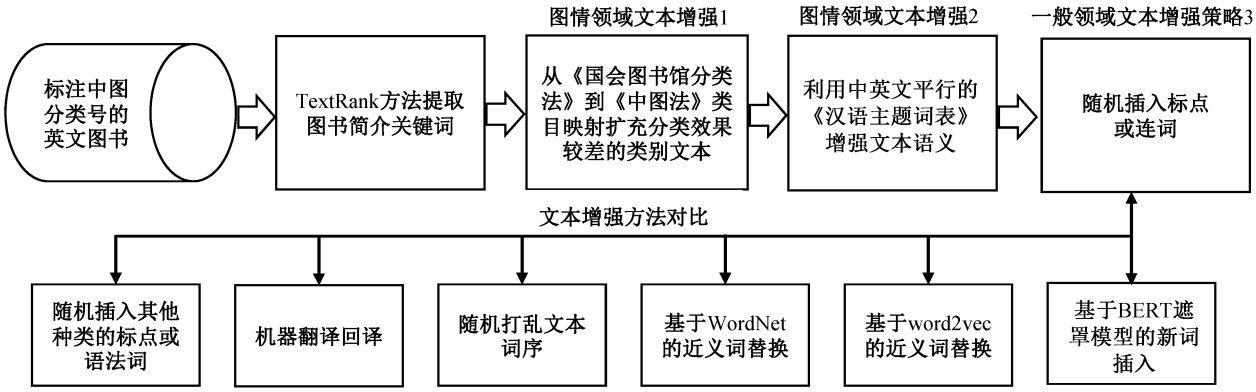

图1　英文图书分类与文本增强策略框架

Fig. 1　Framework of English literature classification and text augmentation

① https://huggingface.co/bert-base-uncased
② https://huggingface.co/allenai/scibert_scivocab_uncased
③ https://huggingface.co/distilbert-base-uncased
④ https://www.nltk.org/api/nltk.stem.porter.html

2.1 图情领域文本增强 1

从《国会图书馆分类法》到《中图法》类目映射(classification mapping), 扩充分类效果较差的类别文本。类目映射是使不同图书文献分类体系关联起来的过程, 通常以不同体系间分类号对应规则的形式表现。如果一册英文图书带有其他体系的分类号, 通过既有的映射规则, 外文图书的其他分类号也能转化为中图分类号。但由于每种分类法层次复杂, 不同的分类法在编制原则、体系侧重点和类目颗粒度等方面存在差异, 所以只能得到粗略的不全面的类目映射结果[29]。另外, 并非所有英文图书都预先标注了其他体系的分类号。因此, 类目映射单一方法稍显力不从心。

我们将类目映射视为文本增强的一种手段, 在得到原始文本分类结果的基础上, 通过类目映射, 扩充分类效果较差的类别的文本。类目映射的源文本采集自“古登堡”网站①。每一个文本都包含图书的标题、关键词和《国会图书分类法》的分类号。映射规则参考蒋彦廷等[30]构建并开放的 106 条中从《国会图书分类法》到《中图法》的单向映射规则, 部分规则如表 2 所示。

通过上述类目映射方法, 我们将古登堡项目网站 19870 册英文图书的《国会图书分类法》分类号转换为中图分类号, 作为文本增强的备用数据。

2.2 图情领域文本增强 2

基于《汉语主题词表》的语义增强。如 1.2 节所述, 《汉语主题词表》(简称《主题词表》)大部分词条由汉语术语、英语翻译和中图分类号组成, 部分术语还涉及多个中图分类号。我们从汉语主题词表服务网站[8]采集各学科领域词条共 11886 个。对于训练集与测试集中的文本, 如果出现上述的英文术语, 就在该文本中补充一个特定的主题词, 表示中图分类号的含义。补充的单词一般是中图分类号一级大类英译的关键词, 例如分类号 C 补充 social, 分类号 D 补充单词 political, 分类号 E 补充单词 military, 分类号 F 补充 economy, 分类号 G 补充 culture, 分类号 H 补充 language, 分类号 I 补充 literature, 分类号 J 补充 art, 分类号 K 补充 history, 分类号 N 补充 natural, 分类号 O 补充 math, 分类号 P 补充 astronomy, 分类号 Q 补充 biology, 分类号 R 补充 medical, 分类号 S 补充 agriculture, 分类号 T 补充 industry, 分类号 U 补充 transport, 分类号 V 补充 aviation, 分类号 X 补充 environment。如果一个术语涉及多个中图分类号一级大类, 则添加多个对应的主题词。

2.3 一般领域文本增强

随机插入标点或连词。前两项依据类目映射、主题词表的增强策略适用于图书情报这一特定领域, 一般领域的文本增强可以推广到其他领域。受 Karimi 等[21]启发, 一般领域文本增强策略的具体步骤如下: 对于单词数为 n 的文本, 随机插入 0.3n (向下取整)个符号, 符号从集合 A={".", ";", "?", ":", "!", ","}或 B={and, or, so, but, as, since}中随机选择。前者的元素均为英文的标点符号, 后者的元素为实义较弱的连词。将随机插入标点符号的文本作为新样本加入训练集, 比较它与如下 6 种文本增强策略的效果。

1) 机器翻译回译: 我们选用基于 transformer 架构的两个机器翻译模型, 分别为 opus-mt-en-zh② (英译中, 1.41 GB)和 opus-mt-zh-en③(中译英, 852 MB)。采

表 2 美国《国会图书分类法》到《中图法》的类目映射表

Table 2 Classification mapping from LCC to CLC

美国《国会图书分类法》类目和含义	映射的《中图法》类目和含义	映射的《中图法》一级类目
TK7885-7895 Computer engineering and hardware	TP3 计算技术、计算机技术	T 工业技术
JZ International law and relations	D8, D99 外交关系, 国际法	D 政治法律
CC Archaeology	K85 文物考古	K 历史地理
GR Folklore	K89 风俗习惯	K 历史地理
HJ Public finance	F81 财政、国家财政	F 经济
HQ Family, marriage, women	C913.1, C913.68 恋爱、家庭、婚姻、妇女	C 社科总论
QL Zoology	Q95 动物学	Q 生物科学

① https://www.gutenberg.org/ebooks/

② https://huggingface.co/Helsinki-NLP/opus-mt-en-zh

③ https://huggingface.co/Helsinki-NLP/opus-mt-zh-en

用“英→中→英”回译路径，生成近似文本。

2) 随机打乱文本词序：每个文本随机打乱词语顺序，合成新文本。

3) 基于word2vec词向量的近义词替换：预训练词向量来自用105 MB图书标题简介语料训练的词向量项目[①]。对于每个单词数为n的文本，随机选中$0.1n$(向上取整)个词语w(除部分连词、介词和冠词等停用词)，利用词向量模型，计算与词语w相似度最高的另一个词语w_1。用词语w_1替换w，生成新文本。

4) 基于WordNet的近义词替换：方法与基于word2vec词向量的近义词替换方法类似，只是在查找近义词时，使用WordNet知识库[②]，从单词w的Synonym set中随机选择一个近义词w_1，用词语w_1替换w，生成新文本。

5) 基于BERT遮罩语言模型的新词随机插入：利用BERT的遮蔽语言模型(masked language model，MLM)机制，对于单词数为n的原文本，随机将每个文本中$0.1n$(向上取整)个词替换为[MASK]符号，使BERT-base-uncased模型完成完形填空任务，预测出可能的候选词。为了不缺损原有信息，将文本还原，并在其末尾插入MLM预测的新词语。若向上取整的$0.1n$大于1，则多次遮蔽原文本的单词，并预测新词。

6) 随机插入其他种类的标点或其他词性的功能词：将上述集合A中的逗号、句号和问号替换为左括号、单引号和双引号。将集合B中的连词替换为助动词、介词、冠词和代词等其他实义较弱的语法词。将其随机插入文本中，生成新的训练样本。

3 实验结果与分析

已标注的中图分类号的英文图书实验数据来自北京师范大学图书馆公开的《外文图书选购目录》。图书领域涵盖从“B哲学”到“X环境、安全科学”共20类。为保证数据平衡，对于图书超过2000册的学科领域，从中随机抽取2000册。对于不足2000册图书的领域，将该领域的所有图书信息纳入实验数据。最终，除V航空航天、U交通运输、N自科总论和E军事4类图书数量分别为684，833，562和1430册(少于2000册)外，其余16类图书数量均为2000册。数据集共包含35509册图书。

① https://github.com/JiangYanting/Pretrained_gensim_word2vec
② https://wordnet.princeton.edu

如1.1节所述，北京师范大学图书馆公开的外文图书选购目录下，每册图书没有标引关键词和主题词。因此如图1所示，我们采用TextRank方法，从简介文本中提取出若干关键词，与书名字段一起作为输入模型的文本。按20%的比例，从35509册文献中划分出测试集7102册。测试集中各类文献数量的比例与训练集一致。在文本增强过程中，我们只扩充训练和验证集，测试集始终不变。

3.1 基于原始数据集的实验

我们将每册文献的标题和关键词作为输入模型的文本。实验所用的GPU为一块RTX 2080 Ti，Cuda版本为10.2。各模型参数设置如下：支持向量机的种类为线性SVM；随机森林的分类树数量上限为200；Fasttext模型词向量维数为300，学习率为0.1，N-gram参数为2-gram，损失函数为Softmax。3种预训练模型的初始学习率均为2×10^{-5}，batch size为32，从训练集中切分出验证集的比例为10%。模型均采用早停策略，训练到损失(loss)在验证集上不再下降为止。测试集上的正确率(Acc)和宏F1(Macro-F1)分数表现如表3所示。

从表3可以发现，首先，无论文本预处理时是否词干化，基于一元语法的Random Forests和SVM的分类效果都比较差，而Fasttext模型在词干化后，Acc与Macro-F1有所提升，但是与BERT等预训练模型相比仍有差距。其次，在3个预训练模型中，BERT-base-uncased均取得最佳效果。压缩蒸馏的DistilBERT虽然模型大小只有BERT-base-uncased的约60%，但其表现与后者相差无几。SCI-BERT虽然曾在114万篇英文论文语料上预训练，但其表

表3　基于原始文献数据的分类实验结果(%)

Table 3　Literature classification based on initial data (%)

分类模型	文本是否词干化	Acc	Macro-F1
Random Forests (1-gram)	未词干化	15.29	17.19
	词干化	16.01	17.65
SVM (1-gram)	未词干化	21.70	23.07
	词干化	21.32	22.24
Fasttext	未词干化	62.98	63.24
	词干化	64.15	64.28
BERT-base-uncased + FCN	未词干化	**71.53**	**71.38**
SCI-BERT + FCN		70.29	70.15
DistilBERT + FCN		71.37	71.22

现不及另外两个预训练模型。我们推测有如下两这方面的原因。第一，SCI-BERT 的预训练论文的分布不平衡。SCI-BERT 的 114 万篇预训练论文，有 18%来自计算机科学，其余 82%来自生物医学领域，缺乏其他领域的语料[31]。在各类的 F1 值表现方面，SCI-BERT 也只有 O 数理类、T 工业技术类和 Q 生物科学类超过 BERT-base-uncased，其余类别的表现皆低于 BERT-base-uncased。第二，用于预训练的论文，其风格与图书数据集中的标题和简介语体不尽相同。基于 BERT-base-uncased 模型分类时，各类的 F1 分数如图 2 所示。

由图 2 可以发现，首先，在数据总规模均为 2000 册的情况下，H 语言文字、S 农业、O 数理科学和化学的分类表现较好。U 交通运输和 E 军事类的图书虽然分别只有 833 和 1430 册，但仍居分类效果前五位。第二，D 政治法律、T 工业技术、C 社科总论和 K 历史地理四类虽然各有 2000 册图书的数据，但分类的 F1 分数均低于 70%，说明它们的图书主题较为广泛和分散，达到相同分类效果需要比其他类别更多的训练数据。第三，N 自科总论的分类效果最不理想，一方面是由于数据量不足造成(N 类图书仅 562 册，另一方面也有该类本身定位和特征的因素：自然科学总论是对具体各类自然科学门类的抽象综合和概述，还涉及科学技术史、系统科学和非线性科学，不可避免地与 O, P, Q 和 X 等具体门类存在千丝万缕的关系，导致分类难度较大。

在文本增强实验中，我们继续使用表现最佳的 BERT-base-uncased 模型。从表 4 可以看出，BERT-base-uncased 模型的效果随图书简介关键词个数的变化而变化。当关键词在 20 个以上时，效果提升不再明显。因此在后续实验中，我们用 TextRank 从每册图书简介里提取最多 20 个关键词，与书名一起作为输入文本。

3.2 类目映射和《主题词表》语义增强的实验

基于图情领域文本增强 1 方法，我们将 19870 册英文图书的《国会图书分类法》分类号转换为中图一级分类号，并从中提取分类效果较弱的 K 历史地理、C 社科总论、T 工业技术、D 政治法律和 Q 生物科学等 10 类共 3465 册英文图书的信息，补充到训练集中。基于图情领域文本增强 2 方法，我们利用中英平行的《主题词表》，搜寻匹配训练集、验证集和测试集中的术语，给术语所在的文本增添《中图法》大类的关键词，在不增添新训练样本的条件下，增强原始数据集中各文本的语义信息。表 5 列出 BERT 在文本增强后的效果优化情况。

由表 5 可知，经由类目映射扩充弱势类以及《主题词表》语义增强后，图书分类的结果均有所上升。专门扩充效果较差的弱势类，宏 F1 值上升较为明显，类别不平衡问题有所缓解。如果直接将 19870 册类目映射后的图书信息全部加入训练集，分类的表现反而下滑。这可能是由于 19870 册图书中，I 文学类占绝大多数(15575 册)，而文学类的分类效果相对较强，大规模扩充强势类别的文本，会加剧数据不平衡，淹没弱势类文本扩充的效果。

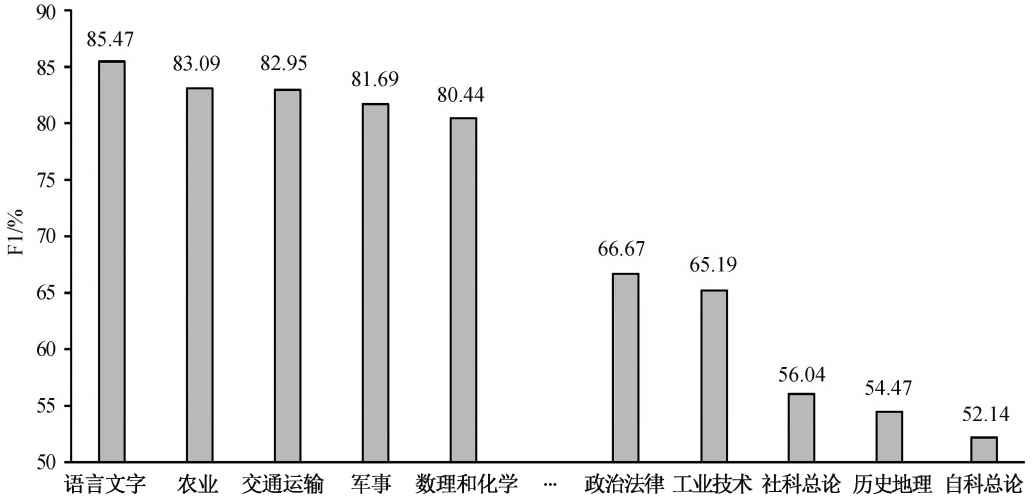

图 2 基于 BERT 的各类别文献分类的 F1 分数

Fig. 2 F1-score of each literature class based on BERT

表 4　基于 BERT 的分类效果随 TextRank 提取的关键词个数变化情况(%)

Table 4　Classification results based on BERT change with num of keywords extracted by TextRank (%)

TextRank 提取的图书简介关键词个数	Acc	Macro-F1
10	71.53	71.38
15	72.16	71.99
20	73.23	73.06
25	73.40	73.32
30	73.51	73.26

3.3 随机插入标点或连词策略与其他策略的对比

我们使用文本随机插入标点方法，给每个文本生成一个新文本，使整个训练集规模扩大一倍，同时比较其余 6 种文本增强方法的效果。实验结果如表 6 所示。

由表 6 可知，第一，在 11 种策略里，随机插入标点的策略 6 表现最佳，正确率与宏 F1 值分别提升 2.14 和 2.34 个百分点，优于机器翻译回译方法、词序随机交换方法、基于 BERT 的 MLM 新词插入方法以及基于 word2vec 或 WordNet 的近义词替换方法。策略 7 中，将插入文本的逗号、句号和问号改为左括号、双引号和单引号后的效果却有所下降。第二，在近义词替换的策略方面，基于 WordNet 的方法优于基于 word2vec 词向量的方法。这里由于 WordNet 作为人工构建的知识库，对近义词的选取比词向量更加严格精准。第三，在随机插入一些意义较虚的语法词的策略方面，随插入词性的不同，总体效果呈现出连词最佳，冠词代词与介词次之，助动词最差的情况。其中，向文本随机插入连词的策略 8 的效果与策略 6 相差无几。

在类目映射扩充弱势类文本的基础上，通过策略 6 随机插入标点和策略 8 随机插入连词，使训练集和验证集文本总数达到 95616，变为原来的 3 倍。最后，查找每个文本存在于《主题词表》的学科术语，增强每个文本的语义。模型在测试集上的正确率和宏 F1 值分别达到 76.84%和 76.41%，比文本增强前(表 4 关键词数目为 20)分别提升 3.61 和 3.35 个百分点。

向原文本随机插入标点或连词的策略较为简便，其表现却超越基于模型、算法、知识库的其他

表 5　类目映射和《主题词表》语义增强后的效果上升幅度

Table 5　Performance gain after data augmentation based on classification mapping and thesaurus semantic aug-mentation

文本增强方法	扩充的训练数据量	Acc 上升的百分点	Macro-F1 上升的百分点
类目映射扩充分类效果较差的 10 类文本(弱势类)	3465	0.21	1.15
类目映射扩充所有类别的文本	19870	–1.20	–0.92
《主题词表》语义增强	0	0.82	0.93
《主题词表》语义增强+类目映射扩充弱势类	3465	0.96	1.21

表 6　一般领域的文本增强策略效果比较

Table 6　Performance comparison among text augmentation strategies on general fields

使训练集规模增加一倍的文本增强策略	Acc 上升的百分点	Macro-F1 上升的百分点
1.　机器翻译回译方法	1.53	1.64
2.　文本词序随机交换	1.72	1.80
3.　基于 word2vec 词向量的近义词替换	1.48	1.68
4.　基于 WordNet 的近义词替换	1.87	2.04
5.　基于 BERT 遮罩语言模型(MLM)的新词随机插入	1.53	1.96
6.　随机插入分号、感叹号、冒号、逗号、句号、问号	**2.14**	**2.34**
7.　随机插入分号、感叹号、冒号、左括号、双引号、单引号	2.01	2.33
8.　随机插入连词 and, or, so, but, as, since	2.11	2.33
9.　随机插入介词 in, on, above, under, for, of	1.42	1.65
10.　随机插入助动词 is, are, was, were, be, have, has	1.27	1.38
11.　随机插入冠词代词 the, this, that, these, a, those	1.48	1.61

文本增强方法，我们认为这与BERT模型中这些字符/词的初始表示有关。本文提取BERT模型顶层的若干词语和字符的768维向量，通过主成分分析降至2维投影至平面，结果如图3所示。

由图3可知，无论是标点符号，还是连词、介词、助动词和冠词等一些语法词，其向量表示与“math”“medical”“military”“geography”等与特定学科关联紧密的主题词界限明显，句号、问号、分号和感叹号等标点与学科主题词的距离尤其远，意味着其向量表示与具体的学科主题无关。基于英文维基百科等海量语料，在BERT完型填空式的预训练阶段，标点和功能词也参与预训练，但由于标点符号和语法词缺乏实义，与它们共现的词语种类众多，分布规律不明显。我们采集909 MB的维基百科语料为样本，统计部分字符与词语相邻的字符/词频次(“相邻”界定在左右各3个词的范围内)，并依据每个字符/词所邻接字符词的频率分布情况，计算其信息熵(information entropy)并降序排列，结果如表7所示。

从表7可以看出，除单双引号外，大多数标点符号、连词和助动词的相邻字词种类和信息熵都高于“economy”“math”“linguistics”等反映特定学科领域的词语。在标点符号方面，逗号、句号和问号的信息熵明显高于左括号、单引号和双引号，从而解

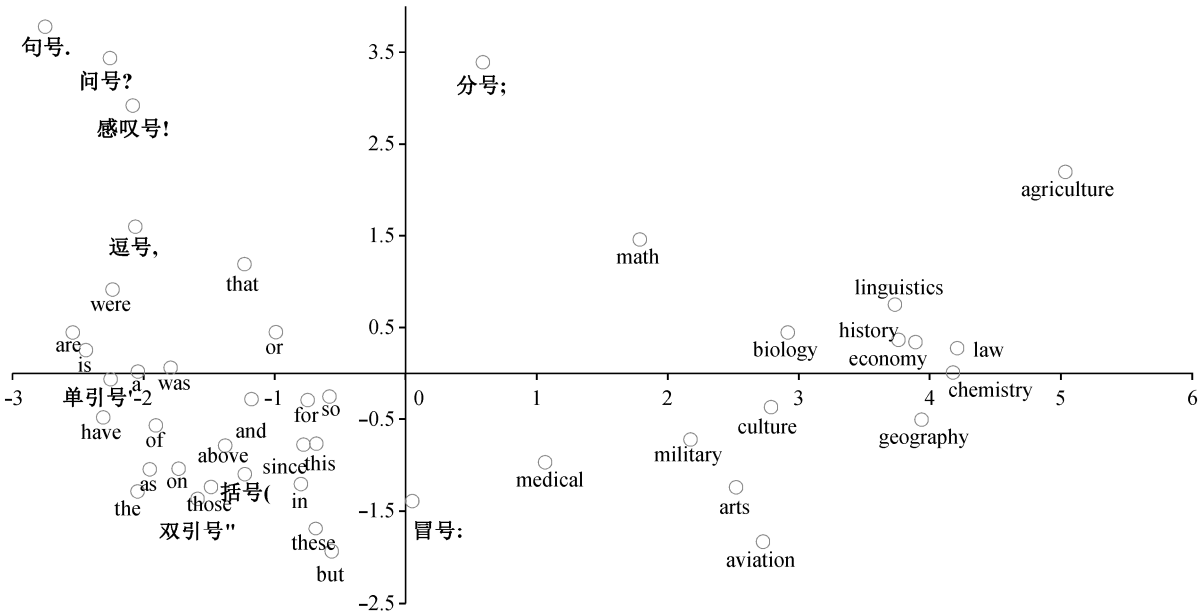

图 3 来自 BERT-base-uncased 顶层的字词向量可视化示意图

Fig. 3 Visualization of word and char embeddings from BERT-base-uncased’s top layer

表 7 部分字词的相邻字词与信息熵统计

Table 7 Statistics of neighbouring words and information entropy of some words

字符/词	相邻字符/词种类	信息熵	字符/词	相邻字符/词种类	信息熵
连词 and	644717	11.184	连词 but	54988	9.929
逗号,	1250162	11.141	左括号(	646696	9.863
句号.	735902	10.871	information	19024	9.776
问号?	30417	10.691	连词 since	28128	9.769
连词 or	123892	10.616	连词 so	27662	9.554
助动词 are	101092	10.486	economy	9846	8.649
助动词 have	58839	10.297	math	2233	8.145
助动词 has	67168	10.134	linguistics	2519	7.882
助动词 is	188111	10.118	单引号'	665818	7.047
连词 as	165707	10.088	双引号"	24722	4.714

释了表6中策略6的效果优于策略7的原因：前三者作为适量的噪声信息，邻接字符/词分布更加复杂，不确定性更强。后三者中的单双引号往往成对出现，且常与“say”等表示说话的单词共现，意味着前三者的向量表示比后三者更加中立，无偏向。

在语法词方面，对比表6中策略8与策略10可知，随机插入连词的策略明显优于随机插入助动词，但在表7中，6个连词的信息熵并不总高于4个助动词。我们认为这可以从语法的角度解释：根据Zhou等[32]对BERT的探针(probing)实验，即使在不微调(fine-tune)参数的情况下，模型在词性标注任务中已能取得超过93%的正确率，十分接近微调的表现。因此，BERT在相当程度上学习了单词的词性和语法信息。如果向文本随机插入助动词、冠词和介词，则文本产生主谓不一致、动词连用、语法角色错误以及搭配不合语法的概率较大。连词的主要功能是在词与词、短语与短语、句子与句子之间起连缀作用，尤其在本文中用TextRank提取了若干关键词的情况下，在关键词之间插入连词对文本原本语法结构的扰动相对较小，造成严重语法错误的可能性较低。另外，连词的相邻字词种类和信息熵指标都不低，表6中插入连词的文本增强效果优于插入其他语法词的现象也在一定程度上得以解释。

4 总结

本文通过对图书馆和文献数据库的实际调研，基于预训练语言模型BERT，结合图书情报(图情)领域与一般领域的文本增强方法，针对面向《中图法》的英文图书自动分类进行探索。首先利用TextRank从图书简介中提取关键词，与书名一起作为输入，然后在BERT文本分类模型下，对比多种文本增强方法，证明了图情领域的类目映射、《主题词表》语义增强与一般领域的标点和连词随机插入策略的有效性。综合上述4种文本增强策略，模型在测试集上的正确率和宏F1值分别提升3.61和3.35个百分点。插入分布情况多样、信息熵较高的标点符号和连词，可在不造成文本语法严重错误的情况下，为文本提供语义均衡的适量的噪声信息，从而防止文本分类模型过拟合，改进模型的表现。

在未来的工作中，我们计划扩大数据集规模，结合更多种类的文本增强方法，以期进一步优化英文图书的中图分类号自动标注效果。

致谢　感谢中国电子科技集团第十研究所提供服务器支持。

参考文献

[1] 中国图书进出口(集团)总公司. 海外图书采选系统[EB/OL]. (2022–07–03) [2022–07–17]. https://www.cnpbook.com/

[2] 北京大学图书馆. 新书通报[EB/OL] (2022–06–29) [2022–07–09]. http://newbooks.lib.pku.edu.cn/index.jsp

[3] 曹晓宽. 如何提高英文图书分类标引的效率. 农业图书情报学刊, 2009, 21(8): 74–78

[4] 中国图书馆分类法编辑委员会. 中国图书馆分类法简本. 5版. 北京: 国家图书馆出版社, 2012

[5] 周沫. 《中图法(第五版)》在西文编目中的应用与发展. 江苏科技信息, 2011(7): 51–53

[6] 李景, 钱平. 叙词表与本体的区别与联系. 中国图书馆学报, 2004, 30(1): 38–41

[7] 中国图书馆分类法编辑委员会. 《中国图书馆分类法》[EB/OL]. (2010–03–17) [2022–07–11]. http://clc.nlc.cn/ztfzfbgk.jsp

[8] 中国科学技术信息研究所. 《汉语主题词表》服务系统[EB/OL]. (2017–01–01) [2022–07–09]. https://ct.istic.ac.cn/site/organize/word

[9] The Library of Congress. Introduction to library of congress subject headings [EB/OL]. (2011–04–26) [2022–07–09]. https://id.loc.gov/authorities/subjects.html

[10] 边钊, 唐娉, 闫珺. 关键词规范化对文献主题信息挖掘的影响——以遥感领域为例. 中国科技期刊研究, 2021, 32(12): 1535–1548

[11] 王昊, 严明, 苏新宁. 基于机器学习的中文书目自动分类研究. 中国图书馆学报, 2010, 36(6): 28–39

[12] 倪斌, 陆晓蕾, 童逸琦, 等. 胶囊神经网络在期刊文本分类中的应用. 南京大学学报(自然科学), 2021, 57(5): 750–756

[13] De Luca E, Fallucchi F, Morelato R. Teaching an algorithm how to catalog a book. Computers, 2021, 10(11): No. 155

[14] 邓三鸿, 傅余洋子, 王昊. 基于LSTM模型的中文图书多标签分类研究. 数据分析与知识发现, 2017, 1(7): 52–60

[15] 蒋彦廷, 胡韧奋. 基于BERT模型的图书表示学习与多标签分类研究. 新世纪图书馆, 2020(9): 38–44

[16] 李湘东, 石健, 孙倩茹, 等. 基于BERT-MLDFA的内容相近类目自动分类研究——以《中图法》E271和E712.51为例. 数字图书馆论坛, 2022(2): 18–25

[17] 张智雄, 赵旸, 刘欢. 构建面向实际应用的科技文献自动分类引擎[J/OL]. 中国图书馆学报, 2022 [2022–08–03]. http://kns.cnki.net/kcms/detail/11.2746.G2.20220624.1437.002.html

[18] Frank E, Paynter G. Predicting library of congress classifications from library of congress subject headings. Journal of the American Society for Information Science and Technology, 2004, 55(3): 214–227

[19] Wei J, Zou K. EDA: easy data augmentation techniques for boosting performance on text classification tasks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Hong Kong, 2019: 6382–6388

[20] Wu X, Lv S, Zang L, et al. Conditional BERT contextual augmentation [EB/OL]. (2018–12–17)[2022–08–03]. https://arxiv.org/abs/1812.06705v1

[21] Karimi A, Rossi L, Prati A. AEDA: an easier data augmentation technique for text classification // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana, 2021: 2748–2754

[22] Luque F M. Atalaya at TASS 2019: data augmentation and robust embeddings for sentiment analysis [EB/OL]. (2019–09–25) [2022–08–03]. https://arxiv.org/abs/1909.11241

[23] 张卫, 王昊, 陈玥彤, 等. 融合迁移学习与文本增强的中文成语隐喻知识识别与关联研究. 数据分析与知识发现, 2022, 6(Z1): 167–183

[24] Ren S, Zhang J, Li L, et al. Text autoaugment: learning compositional augmentation policy for text classification // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021). Punta Cana, 2021: 9029–9043

[25] Zhang B, Sun W, Wan X, et al. PKU paraphrase bank: a sentence-level paraphrase corpus for Chinese // CCF International Conference on Natural Language Processing and Chinese Computing, Dunhuang, 2019: 814–826

[26] Mihalcea R, Tarau P. TextRank: bringing order into text // Proceedings of Empirical Methods in Natural Language Processing. Barcelona, 2004: 404–411

[27] Joulin A, Grave E, Bojanowski P, et al. Bag of tricks for efficient text classification [EB/OL]. (2016–08–09) [2022–08–03]. https://arxiv.org/abs/1607.01759

[28] Schuster M, Nakajima K. Japanese and Korean voice search // 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Kyoto, 2012: 5149–5152

[29] 童刘奕, 张鹏翼. 《中国图书馆分类法》和《美国国会图书馆图书分类法》人工映射分析与差异性探究. 数字图书馆论坛, 2018(3): 53–58

[30] 蒋彦廷, 吴钰洁. 英文文献的《中图法》分类号自动标注研究——基于文本增强与类目映射策略. 数字图书馆论坛, 2022(5): 39–46

[31] Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). HongKong, 2019: 3615–3620

[32] Zhou Y, Srikumar V. A closer look at how fine-tuning changes BERT // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin, 2022: 1046–1061