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

蒋彦廷[1,2] 吴钰洁[2]
(1. 成都航空职业技术学院，成都 610100；2. 北京师范大学文学院，北京 100875)

摘要：给英文文献自动标注《中图法》分类号，能减轻图书馆与文献数据库工作人员的负担，促进跨语言知识检索与中外知识交流。面对既有的标注《中图法》分类号的英文文献数据不足的问题，本文面向预训练语言模型BERT，提出中文文献机器翻译、原始英文文本插入标点或语法词以增强分类模型泛化能力等文本增强策略，以及《美国国会图书馆分类法》到《中图法》的类目映射策略扩充文本数据。实验表明，3种策略均能有效提高文本分类效果。通过上述策略，分类的正确率与宏F1值分别提升约6.1个百分点与7.4个百分点。最后开发并发布了一个小程序，实现给英文文献自动、批量标注《中图法》20类一级分类号的功能。

关键词：预训练语言模型；《中国图书馆分类法》；机器翻译；文本增强；类目映射
中图分类号：G250.2 DOI：10.3772/j.issn.1673-2286.2022.05.007
引文格式：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[J]. 数字图书馆论坛，2022 (5)：39-46.

随着经济社会发展与各领域国际交流日益深化，中国进口外文文献规模不断扩大。在纸质文献方面，根据国家统计局《中国统计年鉴》的数据，2020年中国进口外文图书超过3 200万册[1]。2017—2019年，国家图书馆年均订阅纸质西文文献超过4.9万种。在电子文献方面，截至2019年底，国家图书馆外购数据库中的外文电子图书超过51万种，电子论文超过120万篇[2]。

大量引入外文文献，对图书馆或文献数据库的分类、编目工作提出了较高的要求，也给相关工作人员带来了较重的负担[3]。与中文文献的分类编目相比，加工整理外文文献的难度要更大：一是不同语言、文化之间存在隔阂；二是国内外图书分类体系不同，国内大部分图书馆、电子数据库依据《中国图书馆分类法》（以下简称“《中图法》”）给文献分类。绝大部分中文图书版权页的图书在版编目（Cataloguing In Publication，CIP）数据，都标注了《中图法》分类号（以下简称“中图分类号”）。而许多英文图书依据的是《杜威十进制分类法》（Dewey Decimal Classification）或《美国国会图书馆分类法》（Library of Congress Classification，以下简称“《国会图书馆分类法》”），与《中图法》并未建立直接联系。

给英文文献标注中图分类号，能保持国内图书馆图书管理的一致性，方便读者查阅浏览。故本文提出一种基于预训练语言模型BERT（Bidirectional Encoder Representations from Transformers）与文本增强和类目映射策略的英文文献的中图分类号自动标注方法。

1 相关工作

1.1 国内的英文文献分类情况调研

2022年初笔者调研了全国代表性图书馆、文献数据库网站，展现给读者的英文文献分类方法如表1所示。

《中图法》是树状图书资料分类体系，1975年出版第1版，截至2010年已出版到第5版[4]，包括22个一级类目、250多个二级类以及更多的小类。《国会图书馆分类

法》是美国国会图书馆编制的综合性分类法，包括21个基本大类，每个大类以单个字母作为标记[5]。《中国科学院图书馆图书分类法》（以下简称“《科图法》”），在1958年出版了第1版，采用阿拉伯数字为类目的标记符号，包括25大类和更多的小类。《杜威十进制分类法》将知识分为10个大类，以三位数字代表分类码，截至2004年已出版到第22版。

表1 国内代表性图书馆、文献数据库网站采用的英文文献分类体系

图书馆、文献数据库网站	英文文献采用的分类体系
北京大学图书馆、北京师范大学图书馆、首都图书馆、四川大学图书馆、南开大学图书馆、读秀学术搜索、超星发现系统等	《中图法》
上海图书馆、中山大学图书馆等	《中图法》《国会图书馆分类法》
浙江图书馆等	《中图法》《科图法》
中国科学院文献情报中心等	《中图法》《国会图书馆分类法》《科图法》
国家图书馆、Calis联合目录公共检索系统、武汉大学图书馆、南京图书馆、四川省图书馆、浙江大学图书馆、广东省立中山图书馆等	《中图法》《国会图书馆分类法》《杜威十进制分类法》

调查发现：第一，在文献管理的实际工作中，绝大多数图书馆与文献数据库网站给英文文献分类时，都依据《中图法》；第二，另外有一些机构虽然兼用《中图法》《杜威十进制分类法》《国会图书馆分类法》，但也有主次之分（一方面，给英文图书编制索书号时，仍主要参考《中图法》，而《国会图书馆分类法》与《杜威十进制分类法》的分类号仅在图书数据库中作为次要字段出现；另一方面，这些机构网站中的一部分英文图书仅有中图分类号，而缺失《国会图书馆分类法》分类号）；第三，《科图法》目前在国内图书情报机构中的使用率比《中图法》《国会图书馆分类法》《杜威十进制分类法》低。

笔者认为，国内图书馆、文献数据库主要采用《中图法》给英文文献分类的原因，一是为了保持与中文文献分类的一致性，以我为主，为我所用。中外文文献采用统一的分类号，能提升检索效率，为科学计量提供便利，帮助发现学科新兴热点与学科交叉领域[6]。二是对实体图书馆而言，图书分类号往往是编制索书号的重要基础。国内熟悉《中图法》的读者更多，依据《中图法》编制索书号，也能方便读者查阅文献。上述调研也反映出给英文文献标注中图分类号的必要性。

1.2 分类法类目映射相关研究

类目映射（classification mapping）指的是在不同知识分类体系的分类号之间建立联系的过程。这对外文图书的中图分类号标注也有所裨益。

在映射方法方面，类目映射方法可以分为人工标注与自动映射。人工标注虽然总体上准确率较高，但依赖具体的专业知识，工作量艰巨，标注效率有限[7]。自动映射方法又可以分为4个小类。①基于分类号同现的方法：当同一批图书文献同时标注了两个体系的分类号时，这两个体系的分类号就能建立一定联系[8]。②基于类目相似度的方法：将分类法的每个条目用若干主题词或句子来描述。通过计算不同类目间词句的相似程度，就可以得到两类分类号的匹配度[9]。③基于交叉检索的方法：收集分类法A下面某个分类号a的文献集合，用该文献集合的关键词去检索另一种分类法B表示的文档。统计出检索中分类法B中的高频分类号“b1，b2，b3，…，bn”，就能建立起它们与分类号a之间的关联。但这种映射方法的准确率与覆盖率不高，且往往建立的是一对多的关系[10]。④基于机器学习的类目映射方法。该方法对标注了某个分类号a的文本信息进行训练，得到这个类目的文本二类分类器，然后用该分类器对另一个分类法的类目“b1，b2，b3，…，bn”标识的语料进行分类。分析分类结果，判断类目a与类目“b1，b2，b3，…，bn”之间是否能映射[11]。

在映射的分类体系方面，目前已有学者探索了《中图法》与国际专利分类法（International Patent Classification，IPC）[9,11]，《中图法》与《杜威十进制分类法》[7,12]，《中图法》与《国会图书馆分类法》[13-14]之间的类目映射工作。但由于每种分类法层次复杂，类目众多，加之不同的分类法在编制原则、体系结构、语言文化、类目颗粒度等方面存在差异，相关研究尚无

法给出全面的、精确的类目映射结果。以童刘奕等[14]在教育、心理、数学领域的分析结果为例，从《中图法》到《国会图书馆分类法》建立的799对类目映射关系中，仅有24.5%是完全等同的关系。这意味着无法仅根据类目映射单一方法，给英文文献标注中图分类号。

1.3 基于机器学习的文献分类技术相关研究

文献分类是自然语言处理（Natural Language Processing，NLP）里文本分类技术的子领域。在算法模型方面，BP神经网络、支持向量机[15-16]、决策树（DT）[17]、长短期记忆（LSTM）[18]和BERT模型以及改进的预训练模型[19-20]已被应用到图书、论文的分类任务中。在文献语种与分类号方面，目前已有较多依据《中图法》给中文图书分类的研究[15, 18-19]，以及依据《国会图书馆分类法》[16]《杜威十进制分类法》[17]给英文文献分类的尝试。

总的来看，目前的研究只依据单一体系，给单一语种的文献分类，还没有给英文文献自动标注中图分类号的探索。究其原因，英文图书、论文在出版时并不自带中图分类号，仅在引进中国的图书馆或文献数据库时，才会由相关工作人员标注归类。这导致既有的标注中图分类号的英文文献数据十分稀缺。

1.4 文本数据增强技术相关研究

在文本数据稀疏的情况下，运用文本数据增强（Data Augmentation for Text）技术有助于提高文本分类的效果。文本增强具体包括回译、独立或依赖上下文的词汇替换[20-21]、随机噪声注入[22]、同类文本交叉增强[23]、强化学习[24]等方法。其中基于回译、词汇替换、强化学习的文本增强方法，需要依赖外部的算法、知识库或预训练模型，具有一定成本。另外，依据分类法的文献分类是一个较为特殊的领域，尚未有学者提出专门针对该领域的文本增强方法。

2 英文文献分类与文本增强方法框架

本文的英文文献分类与文本增强方法框架如图1所示。以下将详细介绍文献分类方法与各文本增强方法。

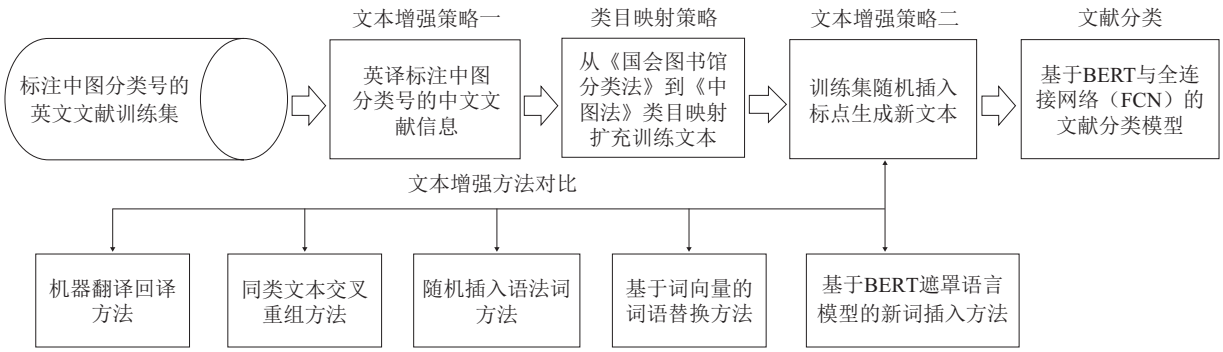

图1 英文文献分类与文本增强方法框架

在文献分类方法方面，笔者基于BERT预训练模型加全连接网络（Fully-Connected Network，FCN）分类器，实现除A类“马列主义，毛泽东思想，邓小平理论”和Z类“综合性图书”之外的中图法一级分类号B到X的20类文献分类。BERT是谷歌2018年发布的基于文本特征提取器Transformer的预训练语言模型，它极大改善了文本语义表示的效果，并在文本分类等各项下游任务中取得了明显突破[25]。一个英文文本输入该模型后，模型提取顶层的符号[CLS]的768维特征向量v作为该文本的向量表示，再后接一个$768\times n$（n为文本类别数量）的全连接层矩阵W，得到一个n维的向量x，最后通过*Softmax*函数归一化，输出文本向量v属于某个类别c的概率$P(c \mid \mathrm{v})$见公式（1），其中*Softmax*函数见公式（2）。

$$P(c \mid \mathrm{v}) = softmax(W \times \mathrm{v}) \quad (1)$$

$$Softmax(\mathrm{x}_c) = \frac{\exp(\mathrm{x}_c)}{\sum_{i=1}^{n} \exp(\mathrm{x}_i)} \quad (2)$$

笔者将比较基于BERT的分类模型与支持向量机（Support Vector Machine，SVM）模型、全连接神经网络模型、Fasttext模型[26]、RoBERTa模型、压缩轻量化的DistilBERT模型的效果。由于BERT等预训练模型在文本预处理时会采用Wordpiece算法[27]，将英文单词

切分为子词，不再需要词干化的预处理步骤。笔者只在文本输入SVM与Fasttext模型前，使用NLTK自然语言处理工具库（nltk.org）将每个单词词干化。

文本增强策略一：如前所述，标注中图分类号的英文文献数据十分稀缺，但标注中图分类号的中文文献（图书、论文等）资源比较丰富。因此笔者尝试采用中文文献英译的方式，扩充英文数据集。采用蒋彦廷等[19]采集整理的中文图书分类数据集，调用百度翻译、阿里云翻译、讯飞翻译的应用程序接口，在保留中图分类号的同时，分别英译中文文献的标题、关键词以扩充数据，并对比基于3种翻译接口的文本增强方法，在文本分类任务上的效果。

类目映射策略：除了将标注中图分类号的中文文献英译的文本增强方法，还可以通过类目映射，将英文图书的《国会图书馆分类法》的分类号转化为中图分类号（见表2）。首先，笔者从古登堡电子书项目网站（gutenberg.org），采集了大量英文图书的标题、关键词与《国会图书馆分类法》的分类号。其次，笔者邀请了两位熟悉外文文献编目的图书馆馆员，请他们依据工作经验，建立了106条映射规则，尽可能将这些书目的《国会图书馆分类法》分类号单向映射到《中图法》上。由于现阶段只针对《中图法》一级分类号进行分类，因此采取“就上不就下”的映射方式：对于《国会图书馆分类法》的类目a与《中图法》的类目b，当人工难以判定a与b是否等同（a=b），且难以判定a是否为b的真子集（a⊆b）时，就让类目a向b的某个上位类c建立映射关系，以保证类目a的含义基本与c等同，或者a的含义能被c囊括。在映射类目颗粒度较粗的情况下，保证映射的稳妥性。

表2 《国会图书馆分类法》到《中图法》的类目映射表（部分）

《国会图书馆分类法》类目与含义	映射的《中图法》类目与含义	映射的《中图法》一级类
QC251 Heat	O55 热学与物质分子运动论	O 数理科学与化学
QD 503 Physical and theoretical chemistry	O64 物理化学（理论化学）	O 数理科学与化学
BF1-990 Psychology	B84 心理学	B 哲学、宗教、心理
LA History of education	G40-09 教育学史、教育思想史	G 文化、科学、教育
TL787-4050 Astronautics，Space travel	V4 航天（宇宙航行）	V 航空航天
GB Physical geography	P9 自然地理学	P 天文学、地球科学
TK7885-7895 Computer engineering and hardware	TP3 计算技术、计算机技术	T 工业技术

通过上述类目映射方法，最终将古登堡电子书项目网站里19 870册英文图书的《国会图书分类法》分类号转换成中图分类号。这批文本数据将添加到训练集中来增强模型的能力。具体实验结果将在3.2节叙述。

文本增强策略二：前两个策略，需要依赖外部的数据集（中文文献数据集、标注《国会图书馆分类法》的英文文献数据集）。而文本增强策略二将不再依赖外部的文献数据，该策略受到Karimi等[22]的启发，具体步骤为：按照对于单词数为n的文本，按30%的比例，在文本中随机插入$0.3n$（向下取整）个的标点符号。标点符号从集合{ “.”, “;”, “?”, “:”, “!”, “,” }中随机选择。随机插入标点符号的文本就作为新的样本，加入训练集中。笔者认为，由于标点符号也参与了BERT模型预训练，存在于模型的词表中，因此在文本分类模型的训练阶段时，向文本插入标点符号，相当于加入了语义均衡的适量噪声信息。这有利于增强模型的泛化能力，从而改进文献分类的效果。

随机插入标点的方法不依赖任何外部数据集与预训练模型，实现十分简易。为验证该方法的有效性，笔者比较其与其他5种文本增强策略的效果。

（1）基于transformer的回译。采用2个基于transformer特征提取器[28]的机器翻译预训练模型，分别为opus-mt-en-zh（英译中，1.41GB，模型地址：huggingface.co/Helsinki-NLP/opus-mt-en-zh）、opus-mt-zh-en（中译英，852MB，模型地址：huggingface.co/Helsinki-NLP/opus-mt-zh-en）。采用“英→中→英”回译路径，给每个文本生成一个语义近似的文本。

（2）同类文本交叉（crossover）重组。每个文本对半切分，同类文本的片段两两交叉，合成新文本。这在保证类别标签基本正确的前提下，改变文本表述合成新样本。

（3）随机插入语法词（grammatical words）。该方法与文本增强策略二随机插入标点类似，只是将随机插入的token集合改为{the，and，of，to，in，on，about，

a}。集合中大都是实义较弱，语法功能更强的词，旨在增强模型的泛化能力，提高模型分类的精度。

（4）基于word2vec词向量的随机换词。选用的预训练词向量模型来自GitHub网站（模型地址：github.com/JiangYanting/Pretrained_gensim_word2vec）。对于每个单词数为n的原始文本，随机选中0.3n（向下取整）个除连词、介词、人称代词、be动词等停用词以外的词语w，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，生成近义的新文本。

（5）基于BERT遮罩语言模型的新词随机插入。利用BERT-base-uncased的遮罩语言模型（Masked Language Model，MLM）[25]。对于每个单词数为n的原始文本，随机将每个文本中0.1n（向上取整）个非停用词替换为[MASK]遮罩符号，用MLM模型预测该符号背后可能的词语。最后为保证原有信息不损失，将文本还原，并在曾被MLM选中的词后面，插入MLM预测的新词语。

3 实验结果与分析

根据Frank等[16]、邓三鸿等[18]对中英文文献的分类经验，每个文本输入的字段为标题和若干反映主题的关键词时，分类效果基本达到最佳水平，摘要字段对文本分类的提升效果不明显。笔者从国家图书馆网站采集了中图法20类，共计36 459册文献的标题与关键词。这些文献绝大部分为图书专著，极少数为论文集。各类文献的数量从高到低依次为：“T工业技术”“F经济”“R医药卫生”“D政治法律”“B哲学宗教心理”“O数理科学与化学”“G文化科学教育”“Q生物科学”“J艺术”“C社科总论”“K历史地理”“I文学”“H语言文字”“P天文地球科学”“S农业科学”“X环境安全”“U交通运输”“V航空航天”“E军事”“N自然科学总论”。

笔者按20%的比例，从36 459册文献中划分出测试集7 292册。测试集中各类文献数量的比例与训练集保持一致。在后续文本增强过程中，只扩充训练数据，测试集始终保持不变。

3.1 基于原始文献数据的分类实验

将每册文献的标题与关键词作为输入模型的文本。各模型的参数设置如下：支持向量机的种类为线性SVM；全连接网络的激活函数为ReLU函数，最大迭代次数为200次；Fasttext模型向量维数为300，学习率参数lr为0.1，ngram参数为2-gram，损失函数为*Softmax*；三种预训练模型初始学习率均为2e-5，每批训练的规模batch size为32，从训练集中切分出验证集的比例为10%。预训练模型均训练到损失在验证集上不再下降为止。测试集上的正确率与宏F1值分数表现如表3所示。

表3 基于原始文献数据的分类实验结果

分类模型	文本是否词干化	正确率/%	宏F1值/%
SVM（1-gram）	未词干化	26.78	22.12
SVM（1-gram）	词干化	27.40	22.14
FCN（1-gram）	未词干化	23.68	21.90
FCN（1-gram）	词干化	24.12	21.53
Fasttext	未词干化	75.55	58.27
Fasttext	词干化	81.25	66.32
BERT-base-uncased	未词干化	84.61	80.85
RoBERTa-base	未词干化	83.04	78.85
DistilBERT模型	未词干化	84.52	80.70

第一，无论文本预处理时是否词干化，基于SVM和FCN的分类效果均不理想，而Fasttext模型在词干化后，正确率与宏F1值分别提升约5.7%与8%。第二，虽然Fasttext模型的正确率接近BERT等预训练模型，但在宏F1值表现上仍比BERT-base-uncased模型低了约14%。这说明BERT模型处理类别不均衡的文本分类任务时，较Fasttext效果更好。第三，在3个预训练模型中，BERT-base-uncased在正确率与宏F1值指标上均取得最佳效果。而DistilBERT模型虽然大小只有BERT-base-uncased模型的约60%，但在分类表现上与后者十分接近。在硬件性能条件有限时，采用DistilBERT模型也不失为良好的折中策略。第四，我们也尝试了文本词干化后再输入预训练模型训练，但分类效果并不及未做词干化时。这也证明预训练模型Wordpiece切分子词方法的良好效果。

统计基于BERT模型分类时，各类别文献的宏F1值表现如图2所示。

第一，虽然如前所述，“T工业技术”“F经济”类的文献数量分别位居第1、2名，但其分类的表现并不在前5之列。第二，虽然“H语言文字”“J艺术”“I文学”“P天文地球科学”类文献数量排名分别位列第13、9、12、14位，但它们的分类表现分别高居第1、2、

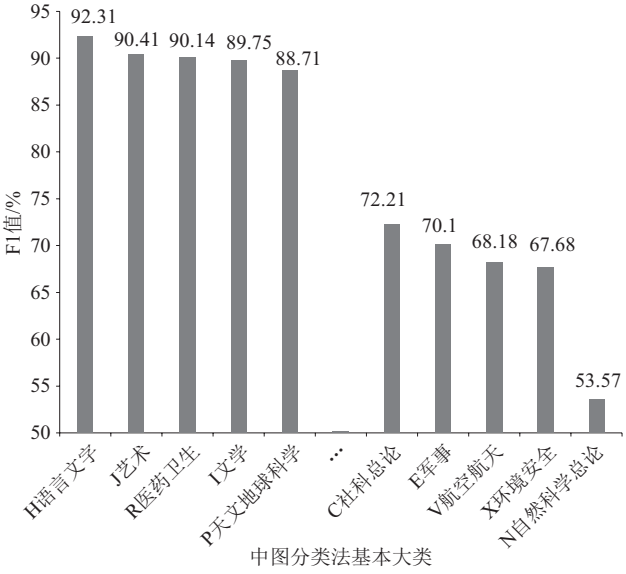

图2 基于BERT的各类别文献分类F1值

4、5位。第三，“E军事”“V航空航天”“X环境安全”和“N自然科学总论”类的文献受限于数据量不足，分类的表现还不太理想。综上所述，T类、F类文献主题较为广泛，自动分类对数据量的要求较高。而H、J、I、P类文献的主题较为集中，提升分类效果所依赖的数据量相对较少。

在后续文本增强实验中，将沿用表3中表现最佳的BERT-base-uncased模型，观察其效果提升情况。

3.2 英译中文文献、类目映射数据增强后的实验

基于第2章所述的文本增强策略一，调用讯飞翻译、阿里云翻译与百度翻译的API，分别将45 570册标注中图分类号的中文图书的标题、关键词翻译成英文，扩充到训练集中。基于第2章所述的类目映射策略，将19 870册英文图书的《国会图书馆分类法》分类号转换成中图一级分类号。表4记录了与原始数据集的分类结果相比，文本增强、类目映射扩充数据后的指标提升情况。

采用讯飞翻译、阿里云翻译和百度翻译英译中文文献，扩充训练集后，文献分类的正确率均有提升。具体而言，讯飞翻译API的效果略优于另外两种，在正确率与宏F1值指标上分别有2.31%与2.95%的提升。这证明了英译中文文献的数据增强策略的有效性。虽然类目映射扩充的数据量不及中文文献英译的方法，但在宏F1值指标上也有0.85%的提升。而将讯飞翻译、类目映射2种方法结合后，模型正确率与宏F1值分别上升2.75%与3.50%，分别达到87.36%与84.35%。一方面，英译中文文献和类目映射能有效改善模型分类的效果；另一方面，效果的提升也反过来证明了机器翻译和类目映射的准确性。

表4 英译中文文献、类目映射策略后的效果上升幅度

文本增强与类目映射策略	训练数据扩充量/册	正确率上升幅度/百分点	宏F1值上升幅度/百分点
讯飞翻译（中→英）	45 570	2.31	2.95
百度翻译（中→英）	45 570	1.66	1.36
阿里云翻译（中→英）	45 570	1.89	2.14
类目映射（国会图书馆分类法→中图法）	19 870	0.46	0.85
讯飞翻译 + 类目映射	65 440	2.75	3.50

3.3 随机插入标点文本增强策略与其他策略的对比实验

在经由文本增强策略一和类目映射，训练集规模达到94 587条的基础上，进而使用第2章所述的文本随机插入标点方法，给每个文本生成一个新文本，从而使整个训练集规模增加一倍。同时，比较了第2章所述的其余5种文本增强方法，具体实验结果如表5所示。

表5 6种使训练数据增加100%的文本增强策略效果比较

使训练集规模增加100%的文本增强策略	正确率/%	宏F1值/%
基于transformer的机器翻译模型回译	87.82	85.21
同类文本交叉重组	88.17	85.55
基于词向量的词语随机替换	89.84	87.80
基于BERT遮罩语言模型（MLM）的新词随机插入	89.96	87.14
原文本随机插入标点	90.69	88.22
原文本随机插入语法词	90.19	87.84

可以看出，基于transformer模型的回译、同类文本交叉重组的2种策略效果较其余策略略差。而随机插入标点的方法有着最佳表现，正确率与宏F1值分别达到90.69%与88.22%，在中图法20类一级分类号分类的任务上，基本达到实用水平。而向原文本随机插入语法词的策略，也有不错的表现，在6种方法里位居第2名。

向原文本插入标点或语法词，不依赖任何预训练

模型或复杂的算法，却表现不俗。我们认为这可能是由BERT模型预训练的方式所决定的：在BERT的遮罩语言模型预训练阶段，标点、语法词参与了预训练，在模型的词表中也能查询到它们的记录。由于标点符号和语法词缺乏实义，与它们相邻的词语分布无明显特征规律。这意味着它们的向量表示不会向任何一类文献的主题偏斜。在模型训练阶段，向文本插入标点符号与语法词，相当于加入了语义均衡的噪声信息，十分有利于增强模型的泛化能力，从而显著提升分类效果。

4　英文文献的中图分类号自动标注小程序设计

笔者汇总了文本增强和类目映射策略扩充的文本数据，在BERT-base-uncased基础上，训练了一个英文文献分类模型，并使用Python语言的tkinter、Pillow与Pyintaller工具库，开发了一个给英文文献批量自动标注《中图法》20类一级分类号的小程序（地址：github.com/JiangYanting/English_books_classification_Program）。用户将每册英文文献的标题与关键词按一册一行的格式写入txt文本文件，上传该txt文件后，系统能在极短时间里，自动标注每册英文文献的中图分类号，并给出预测的概率。预测完毕后，可将预测结果自动保存为txt文件。该小程序界面简洁，使用方便，输出的文件每行各字段之间用制表符分隔，便于存储在Excel、MySQL等结构化数据表中。该程序已初步在某高校图书馆得以应用，有助于提高图书编目、跨语言知识管理与检索的效率，有效减轻文献数据库与图书馆工作人员的负担。

5　总结

给英文文献标注中图分类号是文献知识管理中十分实用、必要的环节，但又面临训练数据不足的问题。本文为基于BERT的文本分类模型提出中文文献的机器翻译方法、《国会图书馆分类法》到《中图法》的类目映射方法、原始英文文本插入标点或语法词以增强分类模型泛化能力的方法。实验表明，3种策略均能有效提高自动分类的效果。向原文本随机插入标点或语法词的数据增强方法简易有效，效果优于原文本回译方法、基于BERT语言模型的完形填空方法、同类别文本拆分重组的方法、基于词向量的近义词替换方法。通过3种策略，分类模型的正确率与宏F1值分别提升约6.1%与7.4%。在未来的工作中，笔者将进一步扩大数据规模，优化模型效果，以实现粒度更细的中图分类号标注功能。

参考文献

[1] 国家统计局. 中国统计年鉴2021 [EB/OL]. [2022-04-01]. http://www.stats.gov.cn/tjsj/ndsj/2021/indexch.htm.

[2] 中国国家图书馆. 信息公开 [EB/OL]. [2022-04-01]. http://www.nlc.cn/dsb_footer/gygt/xxgk/.

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

[4] 国家图书馆《中国图书馆分类法》编辑委员会. 中国图书馆分类法（第五版）[M]. 北京：国家图书馆出版社，2010.

[5] The Library of Congress. Library of Congress Classification Outline [EB/OL]. [2022-04-01]. https://www.loc.gov/catdir/cpso/lcco/.

[6] 蒋彦廷，胡韧奋. 自然语言处理在其他学科领域的影响考察——基于CNKI的中文文献挖掘 [J]. 情报杂志，2021，40（12）：169-176.

[7] 陈瑞，贾君枝. 基于众包模式的分类法映射研究 [J]. 情报理论与实践，2020，43（7）：137-143.

[8] ZHANG Y, JIA P, DI H, et al. Design of Automatic Mapping System between DDC and CLC: Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation [C] // 13th International Conference on Asia-Pacific Digital Libraries. 2011.

[9] 何贤敏，李茂西，何彦青. 基于孪生BERT网络的科技文献类目映射 [J]. 计算机研究与发展，2021，58（8）：1751-1760.

[10] 张建华，李云春，周林志. 基于交叉检索的IPC与CLC映射研究 [C] //中国高等教育学会教育信息化分会第十二次学术年会. 中国高等教育学会教育信息化分会，2014.

[11] 靳雪茹. 基于机器学习的IPC与CLC类目映射方法 [D]. 北京：北京林业大学，2011.

[12] 贾君枝，郝倩倩. DDC到《中图法》类目映射方法研究 [J]. 中国图书馆学报，2013，39（1）：43-50.

[13] 徐烨，肖明. CLC与LCC类目同现映射方法研究——以图情领域为例 [J]. 图书馆论坛，2019，39（12）：11-17.

[14] 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 [J]. 数字图书馆论坛，2018（3）：53-58.

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

[16] FRANK E，PAYNTER G. Predicting library of congress classifications from library of congress subject headings [J]. Journal of the American Society for Information Science and Technology，2004，55（3）：214-227.

[17] DE LUCA E，FALLUCCHI F，MORELATO R. Teaching an algorithm how to catalog a book [J]. Computers，2021，10：155.

[18] 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 [J]. 数据分析与知识发现，2017，1（7）：52-60.

[19] 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 [J]. 新世纪图书馆，2020（9）：38-44.

[20] WEI J，ZOU K. EDA：Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks [C] // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing（EMNLP 2019）. Association for Computational Linguistics，2019.

[21] WU X，LV S，ZANG L，et al. Conditional BERT Contextual Augmentation [C] //19th International Conference of Computational Science. Springer Verlag，2019.

[22] KARIMI A，ROSSI L，PRATI A. AEDA：An Easier Data Augmentation Technique for Text Classification [C] // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing（EMNLP 2021）. Association for Computational Linguistics，2021.

[23] LUQUE F M. Atalaya at TASS 2019：Data Augmentation and Robust Embeddings for Sentiment Analysis [EB/OL]. [2022-04-01]. https://arxiv.org/abs/1909.11241.

[24] REN S，ZHANG J，LI L，et al. Text AutoAugment：Learning Compositional Augmentation Policy for Text Classification [C] // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing（EMNLP 2021）. Association for Computational Linguistics，2021.

[25] DEVLIN J，CHANG M，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（NAACL 2019），Association for Computational Linguistics，2019.

[26] JOULIN A，GRAVE E，BOJANOWSKI P，et al. Bag of Tricks for Efficient Text Classification [EB/OL]. [2022-04-01]. https://arxiv.org/abs/1607.01759.

[27] SCHUSTER M，NAKAJIMA K. Japanese and Korean voice search [C] //2012 IEEE International Conference on Acoustics，Speech and Signal Processing（ICASSP）. Institute of Electrical and Electronics Engineers Inc，2012.

[28] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need [EB/OL]. [2022-04-01]. https://doi.org/10.48550/arXiv.1706.03762.

作者简介

蒋彦廷，男，1997年生，硕士，助教，研究方向：自然语言处理、情报检索，E-mail：jiangyanting@mail.bnu.edu.cn。
吴钰洁，女，2001年生，研究方向：文献学与目录学。

Research on Automatic *Chinese Library Classification* Labeling for English Literature based on Text Data Augmentation and Classification Mapping Strategies

JIANG YanTing[1,2] WU YuJie[2]
(1. Chengdu Aeronautic Polytechnic, Chengdu 610100, P. R. China; 2. School of Chinese Language and Literature, Beijing Normal University, Beijing 100875, P. R. China)

Abstract: Automatic *Chinese Library Classification* labeling can reduce library or literature database staff's burden, promote cross-lingual knowledge retrieval and knowledge communication at home and abroad. Confronting lacking of English literature annotated with *Chinese Library Classification* label, faced with the BERT model, this paper proposes text augmentation strategies which include Chinese literature translating to English and punctuation or grammatical words inserting to improve generalization ability of models. In addition, it proposes the classification mapping from *Library of Congress Classification* to *Chinese Library Classification* to augment text data. Experiments show that these 3 strategies can optimize the performance of text classification. After these strategies, accuracy and Macro F1 score of classification model have respectively increased by 6.1% and 7.4%. Finally, this paper developed and released a programme, which implements automatic and large-batch 20-class *Chinese Library Classification* labeling for English literature.

Keywords: Pre-trained Language Model; *Chinese Library Classification*; Machine Translation; Data Augmentation for Text; Classification Mapping

（收稿日期：2022-04-12）