情报杂志
Journal of Intelligence
ISSN 1002-1965,CN 61-1167/G3

《情报杂志》网络首发论文

题目：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
作者：蒋彦廷，胡韧奋
网络首发日期：2021-09-30
引用格式：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/OL]．情报杂志．
https://kns.cnki.net/kcms/detail/61.1167.g3.20210929.1410.020.html

网络首发时间：2021-09-30 15:36:57
网络首发地址：https://kns.cnki.net/kcms/detail/61.1167.g3.20210929.1410.020.html

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

蒋彦廷[1,2,3]　胡韧奋[2,3]

（1. 四川传媒学院　成都　611745；2. 北京师范大学中文信息处理研究所　北京　100875；
3. 北京师范大学汉语文化学院　北京　100875）

摘　要：[**目的/意义**]探索自然语言处理（Natural Language Processing，NLP）在其他学科领域的影响力，以促进技术的落地应用与创新研究。构建 NLP 主题分类体系与数据集，能为未来相关论文主题识别、NLP 跨学科知识扩散提供有力支撑。[**方法/过程**]利用《中国图书馆分类法》以及论文间的引证关系，从中国知网采集 2 159 篇 NLP 典型文献与 1 376 篇非典型文献，可视化分析文献所属刊物、学科分类号的频次信息，提出 NLP 领域 4 层级主题分类体系，并据此构建论文多主题分类数据集“NLP-others”，进行文献的多标签分类。[**结果/结论**]NLP 在自然、社会与人文各领域均有程度不同的影响，与图书情报学的联系最为密切。相关技术甚至能拓展到处理非自然语言的序列。知识库与知识图谱、神经网络、舆情分析是被广泛提及或应用的技术；LDA、LSTM、CRF、BERT 则是在其他领域应用较多的模型算法。

关键词：自然语言处理；学科交叉；中国图书馆分类法；NLP 主题分类体系；NLP 论文主题分类数据集“NLP-others”；多标签分类

中图分类号：G353.1　　**文献标识码：**A

Influence of NLP on other fields based on data mining of CNKI Chinese papers

Jiang Yanting[1,2,3]　Hu Renfen[2,3]

（1. Sichuan University of Media and Communications, Chengdu, Sichuan　611745;
2. Institute of Chinese Information Processing, Beijing Normal University, Beijing　100875;
3. School of Chinese Language & Culture, Beijing Normal University, Beijing　100875）

Abstract: [**Purpose/significance**] Probe into the influence of Natural Language Processing (NLP) on other fields to promote technology application and research innovation. The NLP topic taxonomy and relevant dataset proposed by this paper can strongly support topic detection of relevant papers and interdisciplinary research of NLP. [**Method/Purpose**] Firstly, making full use of Chinese Library Taxonomy (CLT) and citation network of papers, we collect 2159 typical NLP papers and 1376 atypical NLP papers from CNKI. Secondly, analyze the frequency of journals and CLT labels of papers. Thirdly, we propose a 4-layer topic taxonomy of NLP, and construct a paper multi-topic classification dataset "NLP-others" by this taxonomy. Finally, conduct the multi-topic classification on "NLP-others". [**Result/Conclusion**] NLP has a wide influence on natural science and social science and humanity, and is most closely related to library and information science. Relevant technology can even process sequences besides natural languages. Knowledge base, knowledge graph, neural network and public sentiment analysis are most widely referred or applied technology. LDA、LSTM、CRF、BERT are algorithms or models often applied by other fields.

基金项目：**国家自然科学基金青年项目“面向古籍整理智能化的知识表示与加工研究”（编号：62006021）；教育部人文社科基金项目“国际汉语教材文本可读性智能评价方法”（编号：18YJAZH112）；国家语委“十三五”科研规划重点项目（全球中文联盟专项）“面向国际中文教育的文本可读性智能评价方法研究及分析系统构建”（编号：ZDI135-141）的研究成果之一。**

作者简介：**蒋彦廷**（ORCID：0000-0001-6399-0361），男，1997 年生，硕士研究生，研究方向：自然语言处理、文本挖掘；**胡韧奋**（ORCID：0000-0002-8310-5232），女，1988 年生，讲师，硕士生导师，研究方向：自然语言处理、数字人文。

Key words: natural language processing; Inter-discipline; Chinese Library Taxonomy; NLP topic taxonomy; dataset "NLP-others"; multi-label classification.

0 引 言

近年来,自然语言处理(Natural Language Processing, NLP)作为人工智能的一个分支蓬勃发展。作为一门让计算机有效地理解与处理人类语言的学科,它在文本分类[1]、信息检索[2]、机器翻译[3]、阅读理解[4]等技术上均取得了长足进步。随着社会经济科技发展,信息传播越来越便捷,各个专业学科相互沟通、相互交融的趋势愈加明显。边缘学科乃至跨学科的专著、论文也不断涌现[5]。考察自然语言处理在其他专业的影响,探索 NLP 与其他学科的交叉领域,不仅有利于在学科专业之间找到创新点,助力科学研究;而且也能推动 NLP 技术在各领域应用落地,促进产研合作与研究成果转化。

1 相关研究

在既往的研究中,一些研究者注意到了自然语言处理与其他学科的交叉领域。例如王煜[6]介绍了词频分析、依存句法分析、文本分类、信息检索和知识图谱等技术在建筑工程领域的用途,包括合同管理、工程舆情分析、施工事故原因识别等。薛蕊等[7]指出铁路领域有着大量非结构化文本,NLP 技术能将数据结构化,在铁路智能客服、资产设备管理、智能维修、辅助决策等方面发挥作用。此外还有介绍 NLP 在法律[8]、军事地理情报[9]、教育[10]、社会传播学[11]的应用情况。这一类文献往往是综述性质的,关注 NLP 在某一个具体方面的成果,且考察方式为定性分析。而通过定量方式、尽可能全面展示 NLP 与其他学科领域交融发展、NLP 知识扩散的情况,还是一个值得填补的研究空白。

在运用定量手段发现学科交叉主题、探索跨学科知识扩散的研究中,引文网络、共词分析、聚类法是常用的方法。

引文网络法基于这样假设:引用相似文献的两篇论文,在研究主题上也具有相似性。通过构建共被引网络、进行网络密度、核心度等指标的复杂网络分析,可发现具有相似主题的文献,进而发现学科间主题交叉、知识扩散的现象[12, 13]。采用引文网络进行跨学科的知识扩散探索,主要的关注点是文献之间的引用关系[14],忽略了文献本身的主题内容,主题粒度较粗[15]。该方法难以探寻具体研究主题的跨学科扩散状况。

共词分析法主要以文献关键词为计量项,通过寻找不同学科文献之中共同出现的关键词,构建共现网络,发现交叉研究的主题。共词分析法简便易行,但许多学术文献的关键词设置有很强的人为主观性。一方面部分文献关键词不规范、概念混乱、粒度大小不一[16];另一方面也难以解决多词一义、一词多义的问题,例如论文关键词中的"LSTM"与"LSTM 模型"、"LSTM 网络"、"长短期记忆"、"长短期记忆神经网络"多词一义;而"深度学习"概念在教育学和人工智能领域的含义大相径庭。

聚类法首先通过对不同学科的文献进行语义聚类,构建学科交叉文献集。每一个聚类簇内可能包含学科不同但主题相似的文献。进而对每一个类簇,运用以 LDA(Latent Dirichlet Allocation)为代表的主题模型求解,将交叉文献转变为交叉主题[15]。这是一种无监督数据挖掘的方法,不依赖人为标注的数据。但聚类方法的类别数量往往需要人为设置。在数据量大、样本成员之间总体的语义距离较小的情况下,聚类的效果往往不尽如人意。此外,以 LDA 为代表的主题模型会生成由若干关键词构成的主题,主题的意义需要人为归纳。

我们认为,发现学科交叉研究的主题,包含两个子任务:第一是交叉领域文献集的确定;第二是交叉领域文献集主题的识别。针对这两个任务,该文的研究方法将在第 2 节中详述。

2 研究方法

本文的研究方法涉及两方面。第一,针对交叉领域文献集的确定,我们将利用《中国图书馆分类法》与文献之间的引证关系,确定 NLP 与其他领域的交叉研究文献。第二,针对交叉领域文献集的主题识别,我们构建了一个"数据资源-算法模型-关键技术-应用系统"的 4 层级的 NLP 知识分类体系与文献数据集,通过文献的多标签分类(Multi-label classification),实现有监督的文献主题识别,从而使 NLP 在其他领域的影响作用更具体地呈现出来。

2.1 依据文献分类号、引证关系构建交叉领域文献集

要发现 NLP 在其他学科领域的影响力,首先需要收集 NLP 与其他学科交叉研究的文献。在中国知网(CNKI)论文数据库中,许多论文均标注了文献分类号。文献分类号设置的依据就是《中国图书馆分类法》(以下简称"《中图法》")。《中图法》是一个针对图书、文献的大型知识分类体系,是当今国内图书馆使用最广泛的分类体系。《中图法》包含 22 个一级类别,以下又区分约 250 个二级类别和更多的小类,层层

隶属,逐级细分。

NLP 在《中图法》知识分类体系中的定位是怎样的? 我们经过对代表性 NLP 论文所属分类号的分析调研,认为具有以下《中图法》分类号(以下简称"中图分类号")的文献,就属于 NLP 的典型文献,如表 1 所示:

表 1　自然语言处理的典型中图分类号及其含义

中图分类号	分类号含义
TP391.1	文字信息处理
TP391.2	翻译系统
TP391.3	检索系统
H085	机器翻译
H087	数理语言学
H127	汉字编码

根据《中图法》,TP391.1"文字信息处理"不仅包括文字录入技术,而且也涉及范围更广的、非语音而是书写形式的文字处理系统。TP391.2"翻译系统"与 H085"机器翻译"两个分类号的区别在于,前者主要收录与翻译软件及其应用相关的图书文献,后者则偏重机器翻译及其理论[17]。每篇文献的分类号,由作者或期刊编辑人工标记确定,严谨性和准确性较强。

我们依据上述分类号,从 CNKI 中国知网数据库中收集了 2159 篇文献的题名、刊物名、摘要、关键词、中图分类号等信息。这些包含表 1 典型 NLP 分类号的文献,就是 NLP 领域的典型文献。值得指出的是,这 2159 篇典型文献中,也有不少文献包含了多个中图分类号。

除了采集 NLP 领域的典型文献,我们也收集了 NLP 领域的非典型文献。我们对于 NLP 领域非典型文献的界定标准是:它们虽然本身不含表 1 所示的 NLP 文献分类号,但引用参考了 NLP 领域的典型文献。我们依照此标准,搜寻 NLP 典型文献的引证文献,从中采集了 1376 篇 NLP 的非典型论文。

2.2　建立 NLP 知识分类体系,构建论文主题数据集"NLP-others"　如第 1 节所述,学术论文关键词普遍存在主观性强,一致性较弱的现象。一词多义、多词一义、上位词和下位词的问题也不利于直接通过统计关键词,反映 NLP 与其他学科领域交叉研究的研究主题。另外,根据我们对 3535 篇 NLP 典型与非典型论文的数据统计,论文的关键词同时存在于摘要或标题中的比例不到 30%。这意味着难以通过词向量学习,在论文的标题或摘要中获得关键词的嵌入(embedding)表示。

针对这样的情况,我们依据参考对采集到的部分论文主题的考察,并参考宗成庆[18]对 NLP 领域内容、层次的梳理,构建了一个 4 层级的 NLP 知识分类体系,并据此体系,人工标注了一个 NLP 与其他领域交叉研究的论文主题的多标签分类数据集"NLP-others"。该 NLP 知识分类体系如表 2 所示。

表 2　NLP 的 4 层级知识多标签分类体系

NLP 的层次	该层次下的具体 label
数据资源	语料库、知识库/知识图谱……
模型算法	语言模型、HMM、CRF、最大熵、TextRank、神经网络、词向量、LSTM、SVM、K-means、CNN、序列到序列、LDA 主题模型、TFIDF、注意力机制、BERT……
关键任务	词法分析、命名实体识别、句法分析、语义分析、篇章分析、时间序列分析、网络分析……
应用系统	机器翻译、信息检索、问答对话系统、阅读理解、自动文摘、舆情分析、分类系统、聚类系统、回归系统、生成系统……
其他	其他

NLP 的 4 层级知识多标签分类体系的第一层次是"数据资源",任何一个信息处理系统,都离不开数据和知识库的支持,自然语言处理系统也不例外。第二层次是"模型算法",它主要涉及自然语言处理领域的统计方法与机器学习方法。第三层次是"关键任务",主要涉及从词语、句子序列、篇章等角度,对自然语言文本进行分析并从中提取有价值的信息。第四层次是"应用系统",它是 NLP 知识分类体系中最宏观抽象的一层,其下包含的具体 label,通常都是集成性、实用性较强的落地的系统。

为了让表 2 的分类体系更好地指导 NLP 文献主题数据标注,增强标注的准确度与一致性。我们对该体系做出如下标注说明:

a. 知识库/知识图谱:知识库与知识图谱都属于经由人为提炼、加工后的形式化的知识资源,因此归入同一个 label 中。词典、辞书、本体、语义网、图数据库等主题也归入该 label 中。

b. 语言模型:包括但不限于经典的 n-gram 语言模型与预训练深层语言模型。

c. 神经网络:"神经网络"label 包括"词向量"、"LSTM"、"CNN"、"BERT"等下位概念。当一篇文献中包括这些下位概念时,也需要标注"神经网络"这一上位概念。

d. 词向量、LSTM、CNN、LDA 主题模型等:这些标签分别是所属的一类模型算法的通称。与它们密切相关的改进、变种版算法/模型,也归入对应的标签里。例如"循环神经网络(Recurrent Neural Network)"、"双向的长短期记忆(Bi-LSTM)"模型也归入"LSTM"标签中。

e. 词法分析:该 label 具体包括自动分词、词性标注、词频统计与词语共现相关的内容。另外有关语素(Morpheme)、词类、复合词内部结构等的语言理论研

究,也归入此 label。而“关键词抽取”相关内容不归入此 label,而归入“应用系统”层次的“自动文摘”label。

f. 句法分析:该 label 既涉及短语结构语法、依存语法的自动分析,也包括形式语言、自动机理论、构式语法等语言学语法理论的探索。

g. 语义分析:该 label 主要涉及对语言意义的分析研究,包括语义角色标注、语义依存、词义消歧等。也包括理论语言学领域相关的语义研究(如动词配价理论)。“知识库/知识图谱”label 中涉及语言意义形式化分析的内容(如 WordNet、HowNet 知网),也同时归入“语义分析”label 中。

h. 网络分析:包括图论、复杂网络、社会网络分析等内容。该 label 与“知识图谱”的区别在于,“网络分析”侧重于动态的算法过程与网络性质的分析,例如社群发现、关键节点挖掘、网络表示学习等。

i. 舆情分析:该 label 主要包括监测、情感分析、谣言识别、信息传播等内容。它与网络社交媒体密切相关。

j. 分类系统:该 label 主要包括句子分类、文本分类,也包括广义上的机器学习分类任务。若“舆情分析”label 中涉及到分类任务,也同时标记“分类系统”这个 label。但同层级除了“舆情分析”的其他 label,如命名实体识别、信息检索、问答系统、阅读理解、自动文摘等若涉及了分类的子任务,也不再标记“分类系统”label,以避免类别范围无限制地扩大。

k. 回归系统:该 label 主要涉及对样本数值的预测。例如电影评分预测、温度预测、广告点击率预测、作文评分预测。

l. 其他:当一篇文献不属于其他任何一个 label 时,就标记为“其他”类别。

该分类体系在指导文献主题标注时,以文献的简介信息(包括标题、摘要、关键词)为参考的材料依据。一篇文献可能只有一个 label,也可能有多个 label。在主题标注时,应当彰显文献论述的显式的重点,例如若文献简介明确提到了词向量,除非在文献简介也明确提到了自动分词、词性标注等内容,否则该文献仅标注“词向量”的 label,不标注“词法分析”的 label。

我们依据此分类体系,对采集的文献进行主题标注。在一位 NLP 专业的教师、两位 NLP 专业研究生的合作下,人工标注了每篇 NLP 相关论文的主题 label,构建了 NLP 与其他领域交叉研究的论文主题数据集“NLP-others”。下载链接为:https://www.mediafire.com/file/q5gy8iurtr7am76/NLP_topic_classification_dataset.xlsx/file。它包含 1484 篇带 NLP 主题标记的论文。这 1 484 篇论文或多或少均与其他学科领域有所关联(如表 3 所示)。这为后续的主题统计、论文多主题识别奠定了基础。

3 实验数据

如第 2.1 节所述,实验数据包括 1376 篇 NLP 非典型文献,以及 2159 篇 NLP 的典型文献。而标注的 NLP 论文主题数据集,则包括 1376 篇 NLP 非典型文献的全部,以及 108 篇典型文献。这 108 篇典型文献同时包含了表 1 的 NLP 典型分类号,以及表 1 之外的其他中图分类号。如表 3 所示。

表 3 实验数据的类型及其规模

实验数据类型	数据规模	实验数据的含义
NLP 典型文献	2 159 篇	每篇文献均包含表 1 所示的 NLP 典型分类号
NLP 非典型文献	1 376 篇	每篇文献均参考引用了 NLP 典型文献,但不含表 1 所示的 NLP 典型分类号
NLP 论文主题标注数据集“NLP-others”	1376 + 108 = 1484 篇	NLP 非典型文献数据的全部+108 篇 NLP 典型文献(这 108 篇文献同时包含 NLP 典型分类号与非典型分类号)

4 实验过程与分析

4.1 文献的来源期刊分析 我们首先统计了 NLP 的典型文献、非典型文献来源的期刊分布情况,根据期刊的频次高低绘制了词云图。如图 1、图 2 所示。

对比图 1、图 2 可以发现,NLP 领域的典型文献主要来源于计算机学科相关的学术杂志,尤其以《中文信息学报》《计算机学报》《软件学报》《计算机研究与发展》等为代表。而 NLP 领域的非典型文献,则主要分布在图书馆学、情报学领域的学术期刊中,如《图书情报工作》《情报理论与实践》《情报科学》。这说明,图情领域的许多论文虽然没有标注 NLP 的中图分类号,但参考引用了 NLP 领域的许多典型文献,受到 NLP 的影响较大。

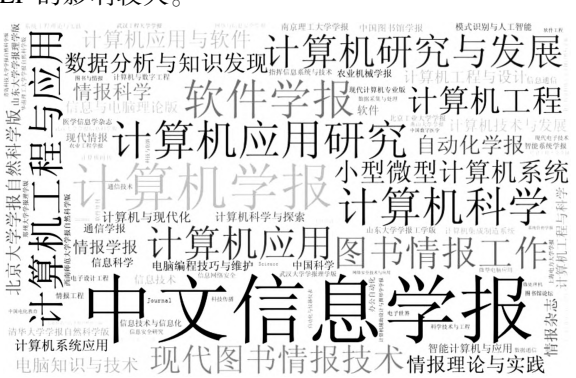

图 1 2159 篇 NLP 典型文献的来源期刊

图 2　1376 篇 NLP 非典型文献的来源期刊

总的来看，除计算机学科外，图情学科与 NLP 的关联最为密切。此外，也可以看到其他领域的学术杂志涉及了一些 NLP 的边缘性、交叉性研究，如医学领域的《医学信息学杂志》、农业领域的《农业机械学报》、教育学领域的《中国远程教育》、传播学领域的《现代传播》、语言学领域的《语言文字应用》、《语言科学》等。

4.2　文献的中图分类号分析　NLP 在其他领域的影响力，可以由 NLP 文献涉及的其他学科分类号的出现频次来定量地衡量。我们统计了 3535 篇 NLP 典型与非典型的文献中，除表 1 以外的其他中图分类号频次。这些中图分类号要么与典型的 NLP 分类号同现，要么是参考引用了 NLP 典型论文的文献的分类号。它们代表着与 NLP 相关的其他学科领域。经归并小类的整理，如表 4 所示。

表 4　与 NLP 相关的其他领域分类号（部分）

中图分类号	分类号含义	分类号出现的频次
TP18	人工智能理论	639
G353	情报资料处理	198
G206	传播理论	143
G252	信息资源服务、文献检索	116
F724、F274	商品流通、企业营销管理与市场	109
G254	信息组织理论	96
G250	图书馆学、情报学工作	85
G434	电化教学、计算机化教学	68

由表 4 可知，与 NLP 有联系的领域十分广泛。限于篇幅，我们阐述分析频次前 8 位的分类号对应的领域。分类号频次最突出的是 TP18 人工智能理论领域，分类号出现的频次高达 639。NLP 作为人工智能的一个分支，与人工智能中的机器学习、知识工程、人工神经网络有着千丝万缕的联系[17]。

频次位居第二是 G353“情报资料处理”的领域。根据对该领域下 198 篇文献的考察，它们主要探讨了各领域知识图谱的构建与应用，以及科研学术信息的挖掘与分析。具体涉及知识图谱[19-21]、主题发现及演化[22,23]等技术。

频次第三的是 G206“传播理论”。该领域涉及舆情管理分析、社交媒体数据挖掘与计算视角下的传播学研究。如唐存琛等[24]通过模块化采集、文本分类与聚类，提升了获取社交网站舆情信息的速度与质量。胡吉明[25]、麻友[26]等分别利用 BiLSTM-CRF、LDA 模型从微博等社交媒体中抽取机构、观点等关键实体，实现舆情的挖掘与结构化。谭振华[27]、刘丽群[28]、徐建民[29]等则从网络传播的角度，对用户转发微博的行为进行特点分析或建模预测。

频次第四的是 G252“信息资源服务、文献检索”领域。如名称所示，该领域着眼于为用户提供有效的信息资源。主要涉及相关数据库、开放数据集的建设[30-31]、知识检索[32-33]、智能推荐[34-36]、问答服务[37]等。

频次第五的是 F724、F274“商品流通、企业营销管理与市场”领域。NLP 在该领域处理的文本类型，既包括电商平台的消费者评论[38-40]，也涉及招聘网站信息[41]与企业微博内容[42]。NLP 发挥的作用主要是挖掘文本关键信息，为企业人员与消费者提供决策支持。

频次第六的是 G254“信息组织理论”领域。该领域的文献主要涉及信息加工、知识标注与结构化工作。例如学术知识描述体系[43]、古籍知识本体[44]、就业知识需求模型的构建[45,46]，也包括机器学习对图书[5]、文献[47]多标签分类相关的研究等等。

频次第七的是 G250“图书馆学，情报学工作”。该领域与 NLP 交叉研究的突出主题，就是图书馆工作的网络化与自动化（数字图书馆）。数字图书馆是未来图书馆的发展趋势，数字人文、文化遗产的数字化[48]以及移动图书馆、数字出版、数字资源的共享[49]都是与 NLP 紧密联系的领域。

频次第八的是 G434“计算机化教学、电化教学”领域。该领域与 NLP 交叉研究的领域较为广泛。包括学生书面成绩的自动评价[50,51]、学习者情感文本分析[52,53]、运用深度学习方法的 MOOC 在线课程信息挖掘[54-56]、知识推荐[57]与教育知识图谱[58]等。

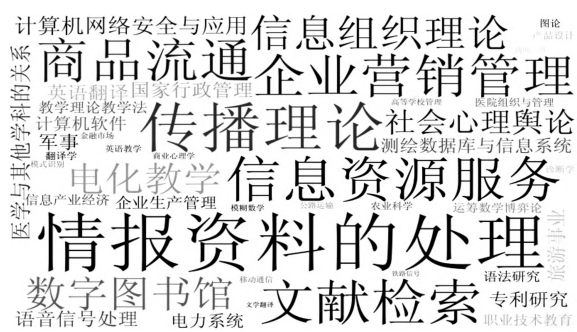

图 3　NLP 联系密切的其他领域一览图

我们按照《中图法》将中图分类号转化成领域名称，根据分类号出现的频次，绘制了图 3 所示的词云

图。可以看出,NLP 与自然、社会与人文学科均有程度不同的影响力。这反映了当代学科间的相互渗透、融合的趋势。只要某领域存在需要处理分析的大量文本,NLP 就能发挥重要作用,例如渔业标准的命名实体识别[59]。此外,NLP 技术甚至在非自然语言的序列中,也有用武之地,例如向量空间模型用于 RNA 序列物种鉴定[60];又如注意力机制、LSTM 模型用于基于实时负荷、历史电价、日期类型、天气等非文本特征的未来电价预测[61]。

4.3 “NLP-others”论文主题数据集的统计与多标签分类 “NLP-others”论文主题数据集标注工作完成后,我们统计了 NLP 主题标签的数量,依据频次绘制了如图 4 的词云,它反映了 NLP 的数据资源、模型算法、关键任务、应用系统被其他学科领域提及或应用的频繁程度。

图 4 NLP 主题在其他学科领域提及/应用的频繁程度

由图 4 可以看出,知识库与知识图谱(占比约 9.71%)、神经网络(占比约 9.23%)、舆情分析(占比约 9.16%)是在其他学科领域广泛提及或应用的 NLP 知识的前 3 位。而篇章分析、阅读理解、自然语言生成、多模态信息处理等目前仍然是富有挑战的 NLP 任务,还有较大的发展空间,因此这些主题在其他领域还较少被提及或应用。在具体模型算法上,擅长序列分析的 LSTM 相关模型、擅长序列标注的 CRF 模型、擅长主题挖掘的 LDA 相关模型、擅长分类的 SVM 模型在其他领域应用较为广泛。此外,近年来兴起的预训练语言模型 BERT 也迅速被与 NLP 相关的其他领域所应用。

为了预测未来产生的 NLP 相关论文的主题,发挥“NLP-others”的主题识别作用,实现知识扩散的精细化探测。我们在“NLP-others”数据集上进行多标签分类。我们选取了 label 数量最高的前 30 个 label 作为多标签分类的标签,其余低频的 label 均转变为“其他”label。

文本多标签分类(Multi-label classification)意味着给每个文本分配一个或多个 label。例如一篇文献同时论述了词向量与语言模型相关的内容,那么它至少应标记表 2 中“词向量”、“语言模型”两个 label。在样本量较少的情况下,它至今仍是一个充满挑战的 NLP 任务[62]。对于多标签分类,目前的常用方法是通过一定手段,将其转化成单标签分类的任务。手段包括二元关联(Binary Relevance)、分类器链(Classifier Chains)、标签子集(Label Powerset)[63]。

二元关联是最简易的方法,对于总计 N 个标签的多标签分类任务,它将训练集中所有属于类别 i 的数据标记为正类,包括多标签的情况,而不属于类别 i 的数据标记为负类别,以此构建 N 个二元分类器。分类器链则在二元关联的基础上,考虑了标签之间的相关性:首个二元分类器只在输入数据上进行训练,之后的分类器则在训练数据和所有之前的分类器上进行训练。标签子集则将数据集中每个多标签的组合情况,都转换为一个单独的类别。从而将原任务转化成单标签的多类分类任务。

我们对所有的文本均按字切分,使用单字、2-gram、3-gram 与 TF-IDF 特征,并把每个 label 的名称在文本中的出现次数作为补充特征。我们选用支持向量机(SVM)与逻辑回归(Logistic Regression,LR)作为分类器。按 9:1 的比例划分训练集与测试集,进行 10 折交叉验证(10-fold Cross-validation)。在测试集上计算每个样本的每个真实 label 的准确率、召回率与 F1 值,如表 5 所示:

表 5 “NLP-others”数据集的多标签分类实验结果(取频次处于前 30 的 label

多标签分类策略	分类方法	F1 值
Binary Relevance	Logistic Regression	67.48%
Classifier Chains	Logistic Regression	67.49%
Label Powerset	Logistic Regression	71.71%
Binary Relevance	SVM	75.57%
Classifier Chains	SVM	76.17%
Label Powerset	SVM	76.60%

如表 5 所示,SVM 分类器显著优于逻辑回归。另外 Label Powerset 的多标签分类策略,略优于 Binary Relevance 与 Classifier Chains,这是因为 Label Powerset 方法把问题转化成单标签多类的分类任务,比起另外两者“一对多”的二元分类,类别间数据不平衡问题有所缓解。30 类多标签分类的 F1 值最优达到 76.60%。这证明对于 NLP 与其他学科交叉研究的文献,“NLP-others”数据集可以成为预测它们论文主题的基础性资源。

5 结 语

该文依据《中图法》文献分类号与文献之间的引证关系,从 CNKI 数据库采集了 3535 篇 NLP 典型与非典型文献。提出了 4 层级的 NLP 知识分类体系,并据此构建了 NLP 论文主题识别数据集“NLP-others”。

实验发现自然语言处理在图书馆学情报学、传播学、企业营销与市场、电化教学、医学信息学、军事学、行政管理、英语翻译、地理信息系统、电力系统等领域均有着广泛的影响。学科领域交叉的态势显著。知识库与知识图谱、神经网络、舆情分析等 NLP 技术在其他学科领域被广泛提及或应用。而篇章分析、阅读理解、自然语言生成等技术在其他领域的应用发展还有较大潜力。我们在“NLP-others”数据集上进行 30 类的论文多标签分类,基于 Label Powerset 方法的 SVM 分类器取得了当前最好效果,F1 值达到 76.60%。实验证明,该文提出的 NLP 主题分类体系,与构建的数据集“NLP-others”能为未来相关论文主题识别、NLP 跨学科研究提供有力支撑。当未来在其他领域出现了引用 NLP 典型论文的文献时,我们不仅可以根据引证关系,将其识别为受 NLP 影响的文献,而且可以利用“NLP-others”数据集与多标签分类算法,识别出 NLP 的哪些具体的数据资源、模型算法、关键任务、应用系统对该领域文献产生了影响。实现知识扩散路径的精细化探测。

由于与 NLP 相关的外文文献大都未标注《中图法》分类号,本研究的数据采集范围限于 CNKI 数据库的中文文献。在未来的研究中,我们将基于外文 NLP 文献数据,探索自然语言处理在其他领域的知识扩散情况。

参 考 文 献

[1] YAO L, MAO C, LUO Y. Graph convolutional networks for text classification[C]//AAAI 2019 Committee. Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA: AAAI Press, 2019: 7370-7377.

[2] 李跃艳, 王 昊, 邓三鸿, 等. 近十年信息检索领域的研究热点与演化趋势研究—— 基于 SIGIR 会议论文的分析[J]. 数据分析与知识发现, 2021, 5(4):13-24.

[3] 冯 洋, 邵晨泽. 神经机器翻译前沿综述[J]. 中文信息学报, 2020,34(7):1-18.

[4] YANG Q, WANG A. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension [C]//ACL 2019 Committee. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL Press, 2019: 2346-2357.

[5] 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.

[6] 王 煜,邓 晖, 李晓瑶,等. 自然语言处理技术在建筑工程中的应用研究综述[J]. 图学学报, 2020,41(4):501-511.

[7] 薛 蕊, 马小宁, 李 平, 等. 自然语言处理关键技术在智能铁路中的应用研究[J]. 铁路计算机应用, 2018,27(10):40-43.

[8] 邱昭继. 文本解析技术及其在法律实践中的应用[J]. 中国法律评论, 2019(2):142-155.

[9] 冯玲莉, 康兴挡. 自然语言处理技术在美军地理空间情报中的应用[J]. 产业与科技论坛, 2016,15(20):50-51.

[10] 王 萌, 俞士汶, 朱学锋. 自然语言处理技术及其教育应用[J]. 数学的实践与认识, 2015,45(20):151-156.

[11] 吴小坤, 赵甜芳. 自然语言处理技术在社会传播学中的应用研究和前景展望[J]. 计算机科学, 2020,47(6):184-193.

[12] Chi R, Young J. The interdisciplinary structure of research on intercultural relations: a co-citation network analysis study[J]. Scientometrics, 2013,96(1):147-171.

[13] 邱均平, 曹 洁. 不同学科间知识扩散规律研究——以图书情报学为例[J]. 情报理论与实践, 2012,35(10):1-5.

[14] 王静静, 叶 鹰. 国际数字人文研究中的跨学科知识扩散探析[J]. 大学图书馆学报, 2021,39(2):45-51.

[15] 阮光册, 夏 磊. 学科间交叉研究主题识别——以图书情报学与教育学为例[J]. 情报科学, 2020,38(12):152-157.

[16] 安兴茹. 我国词频分析法的方法论研究(I)——统计分析要素的界定、分类及问题[J]. 情报杂志, 2016,35(2):75-80.

[17] 国家图书馆中国图书馆分类法编辑委员会. 中国图书馆分类法简本:第五版[M]//北京: 国家图书馆出版社, 2012:207-210.

[18] 宗成庆. 统计自然语言处理:第二版[M]//北京: 清华大学出版社, 2013:51-69.

[19] 卢 恒, 张向先, 尚丽维等. 基于知识图谱的网络社区学术资源深度聚合框架研究[J]. 情报理论与实践, 2021,44(1):180-187.

[20] 王宏宇, 王晓光. 基于大规模开放学术图谱的研究前沿分析框架[J]. 情报理论与实践, 2021,44(1):102-109.

[21] 奥德玛, 杨云飞, 穗志方, 等. 中文医学知识图谱 CMeKG 构建初探[J]. 中文信息学报, 2019,33(10):1-9.

[22] 王效岳, 刘自强, 白如江, 等. 基于基金项目数据的研究前沿主题探测方法[J]. 图书情报工作, 2017,61(13):87-98.

[23] 朱 光, 刘 蕾, 李凤景. 基于 LDA 和 LSTM 模型的研究主题关联与预测研究——以隐私研究为例[J]. 现代情报, 2020,40(8):38-50.

[24] 唐存琛, 王椏可. 一种结合模型集成的舆情管理模型的研究[J]. 计算机应用与软件, 2019,36(6):31-34.

[25] 胡吉明, 郑 翔, 程齐凯,等. 基于 BiLSTM-CRF 的政府微博舆论观点抽取与焦点呈现[J]. 情报理论与实践, 2021,44(1):174-179.

[26] 麻 友, 岳 昆, 张子辰,等. 基于知识图谱和 LDA 模型的社会媒体数据抽取[J]. 华东师范大学学报(自然科学版), 2018(5):183-194.

[27] 谭振华, 时迎成, 石楠翔等. 基于引力学的在线社交网络空间谣言传播分析模型[J]. 计算机研究与发展, 2017,54(11):2586-2599.

[28] 刘丽群, 谢精忠. 结构、风格与内容:社交媒体用户转发的信息特征——基于媒体新冠肺炎疫情报道的考察[J]. 新闻界, 2020(11):39-49.

[29] 徐建民, 韩康康, 何丹丹,等. 融合多种转发习惯的微博转发预测[J]. 情报杂志, 2020,39(03):123-129.

[30] 孙 辉, 王 颖, 张智雄. 基于工具书语料的国史知识库构建和检索[J]. 现代情报, 2016,36(1):64-73.

[31] 张喆昱, 张 磊. 记忆机构的开放数据建设和数字化服务转

型[J]. 图书馆论坛, 2020,40(5):21-26.
[32] 刘爱琴, 安 婷. 面向非相关文献的知识关联检索系统的设计与实现[J]. 现代情报, 2019,39(8):52-58.
[33] 黄孝伦, 王 东, 谭 涛等. 智能科技查新系统的设计与实现[J]. 计算机测量与控制, 2020,28(2):202-205.
[34] 刘海鸥, 孙晶晶, 苏妍嫄等. 面向图书馆大数据知识服务的多情境兴趣推荐方法[J]. 现代情报, 2018,38(06):62-67.
[35] 贾 伟, 刘旭艳, 徐彤阳. 融合用户智能标签与社会化标签的推荐服务[J]. 情报科学, 2019,37(10):120-125.
[36] 蒲姗姗, 何 燕. 个性化学术资源推荐研究:现状、特点及展望[J]. 图书馆学研究, 2019(16):9-17.
[37] 曹树金, 闫欣阳. 社会化问答网站用户健康信息需求的演变研究——以糖尿病为例[J]. 现代情报, 2019,39(6):3-15.
[38] 沈 超, 刘士伟, 徐 滔. 电商平台商家诱导评论的特征与对策研究[J]. 电子商务, 2019(5):47-49.
[39] 毛郁欣, 朱旭东. 面向 B2C 电商网站的消费者评论有用性评价模型研究[J]. 现代情报, 2019,39(8):120-131.
[40] 王忠群, 吴东胜, 蒋 胜等. 一种基于主流特征观点对的评论可信性排序研究[J]. 数据分析与知识发现, 2017,1(10):32-42.
[41] 杨迪月, 葛文博, 黄馨阅, 等. 基于复杂网络的招聘文本挖掘研究——以互联网金融招聘数据为例[J]. 情报探索, 2019(11):75-82.
[42] 夏立新, 张 纯, 陈健瑶等. 企业微博内容对网络口碑及品牌认可度的影响[J]. 情报科学, 2019,37(4):79-85.
[43] 戎军涛. 学术文献内容知识元语义描述模型研究[J]. 情报科学, 2019,37(7):30-35.
[44] 何 琳, 陈雅玲, 孙珂迪. 面向先秦典籍的知识本体构建技术研究[J]. 图书情报工作, 2020,64(7):13-19.
[45] 夏立新, 楚 林, 王忠义等. 基于网络文本挖掘的就业知识需求关系构建[J]. 图书情报知识, 2016(01):94-100.
[46] 俞 琰, 陈 磊, 赵乃瑄. 基于网络招聘文本挖掘的课程知识模型自动构建研究[J]. 图书情报工作, 2019,63(10):134-142.
[47] 马 芳, 黄翠玉. 中文科技期刊论文多标签分类研究[J]. 图书情报导刊, 2019,4(2):26-32.
[48] 李晨晖, 张兴旺, 秦晓珠. 图书馆未来的技术应用与发展——基于近五年 Gartner《十大战略技术趋势》及相关报告的对比分析[J]. 图书与情报, 2017(6):37-47.
[49] 李晓飞. 近五年国内数字图书馆研究可视化分析[J]. 图书馆研究, 2020,50(5):117-128.
[50] 刘 磊, 梁茂成. 英语学习者书面语法错误自动检测研究综述[J]. 中文信息学报, 2018,32(1):1-8.
[51] 吴恩慈, 田俊华. 汉语作文自动评价及其关键技术——来自作文自动评价(AEE)的经验[J]. 教育测量与评价, 2019(8):45-54.
[52] 甄园宜, 郑兰琴. 基于深度神经网络的在线协作学习交互文本分类方法[J]. 现代远程教育研究, 2020,32(3):104-112.
[53] 李 慧. 面向学习体验文本的学习者情感分析模型研究[J]. 远程教育杂志, 2021,39(1):94-103.
[54] 万子云, 陈世伟, 秦斌等. 基于深度学习的 MOOC 作弊行为检测研究[J]. 信息安全学报, 2021,6(1):32-39.
[55] 董庆兴, 李华阳, 曹高辉等. 基于深度学习的 MOOC 论坛探索型对话识别方法研究[J]. 图书情报工作, 2019,63(5):92-99.
[56] 张文德, 李学超, 何 珑. 基于 BiLSTM-CRF 的 MOOC 课程评论抽取研究[J]. 电子设计工程, 2021,29(2):34-37.
[57] 张洁卉, 潘 超, 章 勇. 知识推送在校园网络教学中的应用[J]. 高教发展与评估, 2019,35(5):99-111.
[58] 胡光永. 专业知识与技能体系知识图谱的构建研究[J]. 工业和信息化教育, 2020(12):123-127.
[59] 程 名, 于 红, 冯艳红等. 融合注意力机制和 BiLSTM+CRF 的渔业标准命名实体识别[J]. 大连海洋大学学报, 2020,35(2):296-301.
[60] 祝 斌, 亓合媛, 马俊才. 基于 16S rRNA 序列物种鉴定的改进向量空间模型算法[J]. 计算机系统应用, 2018,27(9):163-169.
[61] 谢 谦, 董立红, 厍向阳. 基于 Attention-GRU 的短期电价预测[J]. 电力系统保护与控制, 2020,48(23):154-160.
[62] 李 锋,杨有龙. 基于标签特征和相关性的多标签分类算法[J]. 计算机工程与应用,2019,55(04):48-55..
[63] Szyma N Ski P, Kajdanowicz T. A scikit-based Python environment for performing multi-label classification[J]. Journal of Machine Learning Research, 2017, 20(6): 1-22.