doi:10.3772/j.issn.2095-915x.2024.06.002

开放科学（资源服务）标识码（OSID）

出版产业链技术关系抽取研究

韦向峰[1,2] 张全[1] 袁毅[1]

1. 中国科学院声学研究所 北京 100190；
2. 富媒体数字出版内容组织与知识服务重点实验室 北京 100038

摘要：[目的 / 意义] 出版产业链中技术与产业链环节的关系对于出版产业技术谱系的构建和出版产业的监测具有重要意义。[方法 / 过程] 设计了传统出版和数字出版的产业链环节，并从业务环节、产业术语、技术术语、参与主体、产品服务等维度进行了产业技术谱系设计。在获取出版产业技术谱系实体后，利用句法依存分析工具获取实体之间的关系模板，使用 Mean Teacher 深度学习训练框架和 BiGRU+Attention 神经网络编码器实现了基于关系模板质量的关系抽取模型；然后使用部分人工标注的半监督深度学习方法对关系模板进行了分类标注和关系分类的模型训练。[局限] 未来仍需研究如何提高关系模板中关系类型的识别准确率，通过改进深度学习模型框架来提高模型的性能。[结果 / 结论] 实验表明该关系抽取模型在实际语料库文本中可获得 66% 的准确率，消融实验表明模板质量等级划分能带来 1% 的正确率提升。

关键词：出版产业；产业链；关系抽取；关系模板；半监督深度学习
中图分类号：G35; G23; TP391

Research on Extracting the Relationship between Technology and Publishing Industry Chain

WEI Xiangfeng[1,2] ZHANG Quan[1] YUAN Yi[1]

1. Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China;
2. The Key Laboratory of Rich-Media Knowledge Organization and Service of Digital Publishing Content, Institute of Scientific and Technical Information of China, Beijing 100038, China

Abstract: [Objective/Significance] The relationship between technology and the nodes of the publishing industry chain is of

基金项目 2023 年富媒体数字出版内容组织与知识服务重点实验室开放基金“基于预训练模型的产业技术谱系构建研究”（ZD2023-11/03）。

作者简介 韦向峰（1976-），博士，副研究员，主要研究方向为人工智能、语音信号与信息处理、大数据与知识组织，E-mail：wxf@mail.ioa.ac.cn；张全（1968-），博士，研究员，主要研究方向为人工智能、语义分析与处理；袁毅（1967-），学士，高级工程师，主要研究方向为计算机软件、计算机应用。

引用格式 韦向峰 , 张全 , 袁毅 . 出版产业链技术关系抽取研究 [J]. 情报工程 , 2024, 10(6): 14-27.

significant importance for constructing the technological spectrum of the publishing industry and monitoring its development. [Methods/Processes] This article designs the industrial chain and the technological spectrum for both traditional publishing and digital publishing. The design of the industrial technological spectrum includes six dimensions, industrial segments, industrial terms, technical terms, participating entities, and product services. After obtaining the entities of the technological spectrum in publishing industry, relationship templates between entities are acquired using syntactic dependency analysis tools. Then, a relationship extraction model based on the quality of relationship templates is implemented using the Mean Teacher deep learning framework and BiGRU+Attention neural network encoder. Furthermore, a semi-supervised deep learning method with partially manually annotated data is employed for relationship classification model training based on relationship template classification. [Limitations] The future research work is still needed on how to improve the accuracy of identifying relationship types in relationship templates and enhance the performance of models by improving deep learning model frameworks. [Results/Conclusions] Experimental results indicate that this model achieves 66% accuracy in actual corpus texts, and categorizing templates can lead to a 1% increase in accuracy.

Keywords: Publishing Industry; Industry Chain; Relationship Extraction; Relationship Template; Semi-Supervised Deep Learning

引言

随着技术的不断发展和进步，出版产业也在向数字化、智能化、创新化发展，并引进和应用了许多新技术，例如人工智能、大数据、云计算、增强现实和虚拟现实等。从出版产业链角度看，传统出版产业的“编—印—发”三个环节[1]（即“编辑、印刷、发行”），可分别对应出版产业链的“上游、中游、下游”，产业链涉及的技术则包括信息采集、印刷工艺、物流技术等；数字出版的产业链可分为“内容提供—数字化加工与出版—内容销售与终端”三个环节，涉及的技术更新更广，例如智能写稿、AI 创作、OCR、XML、语音合成、计算机信息检索、中文信息处理技术、数字加密技术、电子支付、虚拟现实、增强现实等。为了对出版产业上游、中游、下游等链条环节上的技术术语进行组织并构建相应的知识图谱，得到技术在出版产业链中的环节位置、技术关联的专家人才、技术关联的企业信息等知识，需要对出版产业链涉及的业务环节、技术术语、参与主体、产品服务等进行术语抽取和关系抽取，对出版产业涉及的相关技术、产业链环节等构建得到相应的技术谱系，从而在产业技术谱系的基础上实现技术推荐、产品推荐、企业推荐等应用。

本文以《编辑与出版学名词》[1] 为基础，从文本数据中提取了产业链环节、产业术语、技术术语、参与主体、产品服务等产业技术谱系中的术语或实体，并着重研究如何从实际文本数据中获取和完善产业技术谱系中实体之间的关系，具体来说就是要确定技术术语与产业术语之间的关系类型，以及技术术语在产业链中的产业链环节（上游、中游或下游）。这里涉及文本数据的关系抽取，也称为实体关系抽取，是指从给定文本提取出实体对中实体之间语义关系[2]。关系抽取的方法按照是否需要事先抽取出实体，可以分为整体式抽取和级联式

doi:10.3772/j.issn.2095-915x.2024.06.002

抽取[3]。整体式抽取采用统一模型，从非结构化文本中提取得到“实体—关系—实体”三元组。级联式抽取需要先识别实体或已知文本中的实体对，在此基础上从文本中提取得到实体之间的关系。按照是否需要标注语料，关系抽取的方法可分为有监督方法、半监督方法、远程监督方法和迁移学习方法[4-5]。有监督方法需要大量高质量关系标注语料，远程监督和迁移学习方法需要目标领域知识库或者关联领域标注语料，一般用于相似领域的关系抽取，很难应用于特定领域内部的关系抽取。

在本文的研究中，出版产业技术谱系的实体已使用自然语言处理方法提前获取，且获得的出版产业文本语料库没有大量标注好的高质量语料，因此在已有关系抽取研究基础上采用级联式的半监督关系抽取方法[6-8]。本文将重点探讨如何使用半监督的深度学习模型方法，自动或半自动地进行出版产业的技术术语与产业链环节的关系抽取，从而提高出版产业技术谱系的构建效率，为出版产业的技术链监测、产业链监测、创新链监测等提供相关的技术支撑。

1 相关研究

早期的关系抽取采用规则方法来抽取文本中的实体关系，Chinatsu 等[9]利用领域知识专家编写的规则，包括特征词、词性、句法信息和语义模式等，在文本语料库中抽取出与规则匹配的关系实例；Fukumoto 等[10]提出了 OKI 信息抽取系统，其中的关系抽取利用实体之间的谓词来判定实体之间的关系。基于规则的方法主要依赖于专家的领域知识，主观性较大、成本较高，且无法移植到其他领域，对于跨领域的关系抽取构成难以克服的困难，带来高昂的成本。在统计机器学习时期，研究者采用基于特征向量的抽取方法或者基于核函数的方法。特征向量的特征一般包括词语特征、句法特征和语义特征，核函数是隐式地计算特征向量的内积，核函数主要有序列核函数、卷积核函数和树核函数，然后应用于最大熵模型、支持向量机（SVM）、朴素贝叶斯和条件随机场等机器学习模型。2003 年 Zelenko 等[11]首次将核函数应用到关系抽取任务上，使用动态规划的算法，结合两个样本的浅层句法解析树来分析两者间的相似性，采用 SVM 统计机器学习模型在新闻语料文本数据上取得了不错的效果。在中文关系抽取中，车万翔等[12]设计了实体类别、实体位置、前后词信息等特征，采用 Winnow 和 SVM 两种统计机器学习算法，对中文关系抽取标注语料库进行了训练和关系识别，实验表明 SVM 相对于 Winnow 算法可以提高召回率和 F1 值。统计机器学习模型方法在关系抽取中虽然取得了较好的效果，但是该方法依赖于大量标注好的语料库，需要进行大量的预处理工作，耗费的人力成本也很高，且无法自动进行关系类型的扩展。

随着深度学习技术的发展和进步，许多研究者把人工神经网络模型应用到文本数据的关系抽取任务中。2014 年 Zeng 等[13]利用深度的卷积神经网络（CNN）进行特征提取，使用词向量和卷积提取文本中词语和语句的特征，然后通过分类器可以预测两个标记实体之间的关系。Socher 等[14]最先将循环神经网络（RNN）模型用于关系抽取中，文本中每个单词均由向

量和矩阵表示单词本身语义和对其他单词的修饰作用，这种表示可以自动学习到较长短语的深层语义，但模型需要学习的参数过多。长短期记忆网络（LSTM）、门控循环单元网络（GRU）和图卷积神经网络（GCN）等人工神经网络模型也纷纷被引入到文本的关系抽取中，促进了关系抽取的研究，成为目前的研究热点之一。Transformer模型提出了注意力机制（Attention），融合了语义的分布式表示和长距离语义关系发现，从而克服了RNN计算成本高、没有长距离文本语义处理的缺陷。2018年BERT预训练语言模型在Transformer的基础上应用掩码语言模型、预训练和微调技术，在自然语言处理相关任务中取得了很大进步。预训练语言模型BERT自然也被引入到实体抽取和关系抽取的任务中，例如马江微等[15]使用BERT模型作为输入文本的编码器，采用分层强化学习方法分别进行关系与其对应实体的解码，在公开的NYT10、NYT10-sub数据集上F1值分别达到71.8%和69.0%。近年来，以GPT（生成式预训练Transformer）为基石的大语言模型获得了空前发展，其特点是基于海量数据进行了模型的预训练，结合指令学习、人类反馈强化学习、思维链和模型微调等技术，可以根据输入的自然语言指令文本生成指令相关的结果文本，也可以作为人工神经网络模型编码器实现文本关系抽取任务。但是，人工神经网络模型和预训练语言模型用于关系抽取都需要大规模标注的带关系标签的语料库。大语言模型虽然不需要标注语料库，但是其效果依赖于预先训练的文本数据。对于没有标注过的行业领域文本数据的关系抽取，以及大语言模型未进行微调或特殊预处理的专业领域文本中的信息抽取，无法直接应用人工神经网络模型、预训练语言模型或大语言模型来进行文本中关系的抽取。因此，一些研究者转向研究如何利用半监督的方法，只需要少量的人工参与和标注语料，实现半自动地从文本中提取出实体之间的关系。

半监督关系抽取方法只利用少量标注数据，结合相关算法学习训练大量未标记的文本语料，从而抽取得到文本语料中的关系。这不仅可以有效减少人工参与以及对大规模标注语料的依赖，而且关系抽取的性能也能得到提升，还可以扩展应用到其他领域的大规模文本语料的关系抽取中。半监督关系抽取方法的难点主要是初始种子关系的质量问题和如何降低迭代过程中的噪声问题，且少量的人工标注数据容易产生语义漂移。解决问题的办法之一是提高种子关系的质量，或者通过人工总结全面的种子关系。Mean Teacher模型[16]是一种被广泛地应用于半监督的图像分类和语义分割任务的半监督分类模型，解决了时间集成（Time Ensembling）模型在大数据集下更新缓慢和无法实现在线训练的问题。Mean Teacher半监督分类模型的核心思想是：模型既可以作为教师模型，也可以作为学生模型。当作为教师模型时，可以用来产生学生模型学习时的目标，教师模型的参数由历史上前几个学生模型的参数经过加权平均得到；当作为学生模型时，则利用教师模型产生的目标来进行学习训练。与时间集成模型相比，无标签数据的目标标签来自教师模型的预测结果，且教师模型参数无需等到一个epoch训练结束后才更新。模型在CIFAR-10数据集上的实验把当时的SOTA结果错误率从

doi:10.3772/j.issn.2095-915x.2024.06.002

10.55% 降到 6.28%，在 ImageNet2012 数据集上把错误率从当时 SOTA 结果的 35.24% 降到了 9.11%。

本文研究的出版产业链领域的实体以及实体之间的关系并没有大量标注好的语料库，因此需要在产业链技术谱系设计的基础上抽取出出版产业链中的各种实体，同时通过语料库总结归纳出常见的关系类型。关系抽取的实质就是对实体之间的关系类型进行自动分类，本文的具体方法是首先通过依存句法分析得到实体关系模板，然后通过人工标注部分少量的高质量关系模板，对得到的剩下的关系模板再按照与高质量关系模板的相似度分为两种模板（即中等质量关系模板和低质量关系模板）。最后，利用 Mean Teacher 半监督分类模型，把高质量关系模板训练得到的分类模型作为教师模型和学生模型，把中等质量关系模板和低质量关系模板在教师模型指导下训练得到学生模型，最后综合教师模型和学生模型得到一个新的分类模型，用于语料库文本中的关系抽取，也就是关系类型分类。

2 出版产业技术谱系设计

产业技术谱系的设计离不开产业链的分析及其相关概念。产业链是产业经济学中的一个概念，其概念起源于社会分工，产业链的内涵包含了技术链、供需链、价值链、产品链和空间分布组织等方面的链条，不同视角研究会形成不同的产业链环节。例如，郁义鸿 [17] 将其定义为从原材料或矿产资源出发，经过加工、生产直至最终产品达到消费者手中的整个纵向链条；芮明杰等 [18] 将其理解为生产最终产品和服务所经历的从原材料到最终消费品的各个阶段，既涵盖厂商内部过程，也包括厂商间交易的各环节；魏后凯 [19] 则将其视为基于分工经济的一种产业组织形式，强调了供应商、制造商、分销商以及零售商等加盟节点企业之间的分工合作关系；郑大庆等 [20] 将其定义为产业内部企业或部门之间基于技术经济关联和逻辑、空间分布关系形成的一种关联模式。根据距离产业最终产品的远近程度，产业链可以分为上游、中游和下游三个产业链环节。针对传统出版产业，普遍观点认为传统出版产业链是“编—印—发”的链式结构（分别对应上游、中游、下游）；针对数字出版产业，本文借鉴邓佳佳 [21] 的观点（上游为内容提供商、中游为内容出版商和技术服务商、下游为内容销售商和终端设备商），将数字出版产业链分为“内容提供—数字化加工与出版—内容销售与终端”。本文以产业链环节（上游、中游、下游）为纵轴，以业务环节、产业术语、技术术语、参与主体、产品服务为横轴，设计构建了出版产业的技术谱系知识，如表 1、表 2 所示。

在获得类似于表 1 和表 2 的出版产业的相关知识谱系实体之后，进一步可以挖掘实体之间的关系，得到诸如技术术语与产业术语、参与主体、产品服务的关系之后，可构建得到出版产业的技术谱系知识。为了获得出版产业技术谱系中的实体（主要是产业术语、技术术语、参与主体、产品服务），本文主要以《编辑与出版学名词》[1]、出版技术相关论文、相关企业信息为基础，使用自然语言处理的关键词分析技术作为工具，从非结构化文本中获取实体候

表 1　传统出版产业技术谱系设计

产业链环节	业务环节	产业术语	技术术语	参与主体	产品服务
上游	编辑	例如：选题、组稿、审稿、校对、装帧、排版	例如：信息采集、信息处理	例如：责任编辑、总编辑、出版社、杂志社、报社	例如：书稿、文稿、校对稿、送审稿、完成稿
中游	印刷	例如：模拟打样、图文合一、投影制版、印刷材料、喷墨印刷机、数字印刷机	例如：激光成像、激光照排、静电喷墨、离子成像、直接热成像、油墨乳化	例如：印刷厂、印刷复制单位	例如：样书、印刷品
下游	发行与经营	例如：本埠发行、外埠发行、总发行、编发合一、集中征订、电话征订、订阅代理	例如：电子支付、移动支付	例如：新华书店、发行人、总发行单位、报刊亭、读者、书商、发行代理商	例如：图书、报纸、期刊、音像制品、电子出版物

表 2　数字出版产业技术谱系设计

产业链环节	业务环节	产业术语	技术术语	参与主体	产品服务
上游	内容提供	例如：数字内容、数字资源、数字信息	例如：智能写稿、智能校对、自动化排版、语音识别、中文信息处理、人工智能	例如：出版社、杂志社、报社、内容原创网站、游戏开发商、手机内容提供商	例如：原创网文、动漫、游戏、彩铃、彩信、手机报、传统出版物的纸质内容
中游	数字化加工与出版	例如：内容加工、数字化加工、知识化加工、数据清洗、数字内容重组、移动出版、数字出版、云出版、网络出版、跨媒体出版、按需出版	例如：语音识别、语音合成、移动互联网、云计算、大数据、数字签名、虚拟现实、增强现实、数字版权保护技术、中文信息处理	例如：数字音频编辑、视频剪辑师、互联网服务提供者、互联网内容提供者、数字技术提供商、新媒体制作中心、增值服务提供商、信息服务商	例如：数字音像、数字视频、电子书、电子出版系统、内容管理系统、信息检索系统、数字出版平台、数字出版产品、网络出版支撑系统
下游	内容销售与终端	例如：数字产品分销、数字营销、电子阅报栏、报网互动、移动终端、多媒体终端、智能手机	例如：电子支付、移动支付、图形用户界面、多媒体、全媒体、富媒体	例如：订阅服务提供商、数字内容集成商、数字内容分销商、信息服务商、数字内容零售商、终端设备制造商	例如：电子出版物、网络出版物、电子期刊、数字动漫、科学网络数据库、全文数据库、手机游戏、Kindle、iPad、电子书阅读器、阅读设备

选，然后获取出版产业技术谱系中的实体。《编辑与出版学名词》的内容包括综论、编辑、印刷、音像复制、发行与经营、数字出版、出版物、著作权等 8 部分，共 3380 条名词术语，每条名词术语均给出了定义或注释，这些术语和定义注释文本可以作为实体抽取的语料库。以“出版”“数字出版”“产业”“技术”为关键词组合，检索了万方数据知识服务平台，经过筛选后得到了发表年份为 2013—2023 年的论文共计 1528 篇，把这些论文的标题、摘要、关键词也加入到语料库中。此外，从顺企网搜索“出版”相关的企业，得到出版相关的共 2000 家企业的“主要产品”和“经营范围”等描述文本，也作为非结构化的文本加入到语料库中。在非结构化语料库文本的基础上，使用无监督术语抽取的方法获得术语候选。无监督术语抽取方法的步骤如下：（1）按语句标点、回车换行等特殊字符将文本切分为语句文本；（2）对

doi:10.3772/j.issn.2095-915x.2024.06.002

语句文本使用 Python 中的 Jieba 分词模块进行分词处理，得到语句文本中的词语；（3）计算词语的词频、N-gram、TF-IDF、NC-Value[22]、PMI[23] 等统计特征，对统计特征进行加权平均后得到一个总分数；（4）将总分数降序排列得到所需数目的术语候选集合。对于术语候选集合，经过人工筛选审核后得到出版产业技术谱系的术语实体，以《编辑与出版学名词》为准，补充和完善所得到的术语实体，并按照表 1 和表 2 所示分为产业术语、技术术语、参与主体、产品服务等类别。本文对非结构化文本进行处理后得到术语实体候选集合约 2500 个，经过人工筛选和补充完善后得到术语实体 997 个，其中产业术语 491 个，技术术语 193 个，参与主体 105 个，产品服务 208 个。

3 关系抽取方法

本文中出版产业技术谱系的实体已使用自然语言处理方法提前获取，且本文获得的出版产业的文本语料库没有大量标注好的高质量语料，因此在已有关系抽取研究基础上采用级联式的半监督关系抽取方法。具体的抽取方法是：第一，定义关系模板，将产业技术谱系中的实体对（例如：技术术语—关系—产业术语）在文本语句中所对应的关键词序列定义为关系模板，利用自然语言处理的依存句法分析等工具，可以自动提取得到出现实体对的语句文本所对应的实体对之间的关系模板；第二，对关系模板划分等级，分为高质量、中等质量和低质量三类，高质量关系模板为根据出现频率人工筛选并标注后得到的关系模板，中等质量关系模板为与高质量关系模板语义相似度高于一定阈值的关系模板，低质量关系模板为与高质量关系模板语义相似度低于一定阈值的关系模板；第三，对关系模板中的文本词语，利用词向量工具得到词语的词向量表示，将高质量关系模板的词向量表示和关系模板的关系类型（即分类结果）输入到编码器（分类器）中进行深度学习模型训练，得到一个教师模型分类器；第四，使用教师模型分类器对中等质量和低质量关系模板进行预测得到伪标签（分类结果），然后利用伪标签训练分别训练得到中等质量关系模板和低质量关系模板对应的分类器模型；第五，将高质量、中等质量、低质量关系模板对应的分类器模型参数进行加权平均，得到一个总的模板关系分类模型。使用这个训练好的分类模型就可以预测技术术语与产业术语之间的关系类别。

3.1 关系模板和等级

本文的关系模板定义为文本语句中包含实体对及其关系描述的关键词语序列，借助自然语言处理的依存句法分析工具（如 HanLP[24]）可以自动从语句中提取得到关系模板。如语句“人工智能技术可以帮助编辑人员更高效地生成和编辑内容”经过依存句法分析后得到如图 1 所示的结果。

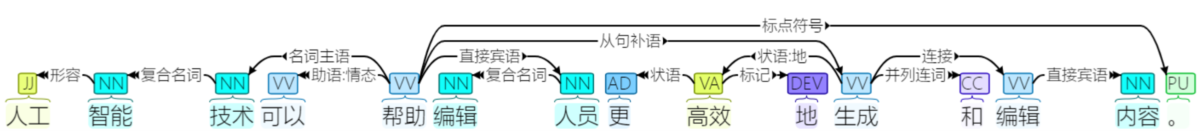

图 1　依存句法分析结果例子

其中得到的包含头实体和尾实体的最短依存路径（Shortest Dependency Path, SDP）为：“头实体 帮助 生成 尾实体 内容”，该 SDP 就是包含头实体“人工智能技术”和尾实体“编辑”的关系模板，其关系类型为“提升”即“技术术语—有助于—产业术语”，而产业术语“编辑”所属的业务环节为传统出版产业链的上游，因此也得到了人工智能技术与产业链上游环节的关联关系。

虽然可以自动得到包含头实体和尾实体的语句所对应的关系模板，但无法自动获得头实体和尾实体之间的关系类型（如上述例子中的关系类型“提升”）。为此先人工标注一部分关系模板的关系类型，获得高质量的关系模板。由于标注成本限制且数量较少，这部分关系模板不能直接用于人工神经网络分类器的训练和学习。对于未经人工标注的关系模板，根据该关系模板与高质量关系模板的质量相似度分为两类不同的模板，以使用不同的训练策略进行分类器模型训练，同时可避免低质量模板带来的噪声和错误。质量相似度包括语义相似度、置信度和语义丰富度。

为了计算未标注模板 p_u 与已知关系类别的高质量模板的语义相似度，可以利用词向量工具（如 Word2Vec[25]），计算两个模板（由词语及其向量构成）的向量欧式距离。每一个关系类别的高质量模板可能会有很多个，为保证计算结果的可靠性同时减少计算耗时，取前 K 个最相似的语义相似度得分的平均值作为未标注模板与关系类别 m 的语义相似度。即首先获得构成模板的词语组合向量，计算未标注模板 p_u 与某个关系类别 m 下所有高质量关系模板的语义相似度，然后得到 K 个最相似的语义相似度得分的平均值作为语义相似度 ClassSim$(p_u,P_{H,m})$。其中，$P_{H,m}$ 表示高质量关系模板中属于关系类别 m 的模板集合。

假设总共有 M 个关系类别，为了突出未标注模板与最相似的关系类别的相似程度，设计了一个置信度指标 $conf_u$（见式（1））。因为语义相似度的取值为 0 到 1 之间，语义相似度越大，那么 1 减去语义相似度的值就越小，取 log（自然对数）之后再取负值则越大，因此语义相似度越大则置信度指标的值也越大。同时需要对不是最相似的关系类别做消除相似的计算，置信度通过式（1）计算得到。

$$conf_u = -log(1-\mathrm{ClassSim}(p_u, P_{H,1_{st}})) - \sum_{k\in\{2,\dots,M\}} log(1-\mathrm{ClassSim}(p_u, P_{H,1_{st}}) + \mathrm{ClassSim}(p_u, P_{H,k_{st}})) \tag{1}$$

在式（1）中，$conf_u$ 为计算得到的置信度指标，$\mathrm{ClassSim}\left(p_u, P_{H,k_{st}}\right)$ 则是未标注模板 p_u 与所有类别中排在第 k 位的关系类别（高质量模板）的语义相似度，计算方法如前文所述。

语义丰富度主要考虑关系模板的长度，可以取长度的三角正切函数的倒数，使得其数值处于 0 到 1 之间。因为三角正切函数为单调递增函数，所以取倒数和负值后仍然是单调递增函数，即模板长度越长，其语义丰富度越大。语义丰富度根据式（2）计算得到。

$$dive_u = -\tan^{-1} Len(p_u) \tag{2}$$

在式（2）中，$Len(p_u)$ 是关系模板 p_u 的长度，tan 是三角正切函数，$dive_u$ 是语义丰富度。

最后，综合语义相似度和语义丰富度，得到未标注关系模板 p_u 的质量相似度得分值，其值也在（0, 1）区间。质量相似度得分 $qual_u$ 的

doi:10.3772/j.issn.2095-915x.2024.06.002

计算方法如式（3）所示。

$$qual_u = dive_u \times \log(1 - conf_u) \quad (3)$$

在式（3）中，$conf_u$ 是关系模板 p_u 根据式（1）计算得到的语义相似度，$dive_u$ 是根据式（2）计算得到的语义丰富度。$qual_u$ 则是未标注关系模板 p_u 的质量相似度得分。

计算得到所有未标注关系模板的质量相似度得分后，依据数据和经验设定阈值 τ，质量相似度得分高于阈值的未标注关系模板归入中等质量关系模板集合，质量相似度得分低于阈值的未标注关系模板归入低质量关系模板集合。由此，所有关系模板被划分为三个等级：高质量关系模板、中等质量关系模板、低质量关系模板，在训练人工神经网络分类器模型时对不同质量关系模板采取不同的训练策略（损失函数），不同程度地利用不同等级关系模板的类别标签及其语义信息，实现从自然语言文本语句到关系模板，再从关系模板到关系类型标注的关系分类或抽取。

3.2 分类器模型和训练

在人工神经网络分类器模型中，分类器的作用是接收输入的向量数据，输出数据的分类类别。常见的人工神经网络分类器模型有循环神经网络（Recurrent Neural Network, RNN）、长短期记忆网络（Long Short-Term Memory, LSTM）和门控循环单元（Gate Recurrent Unit, GRU）模型。深度学习的注意力机制（Attention）借鉴了人类视觉注意力机制，其本质是从众多信息中挑选出对于当前任务更加有用的信息，这不仅提升了模型处理数据的效率，同时保证了预测结果的准确性。因此本文采用双向门控循环单元加注意力机制（BiGRU+Attention）的神经网络架构作为深度学习模型中的分类器模型。

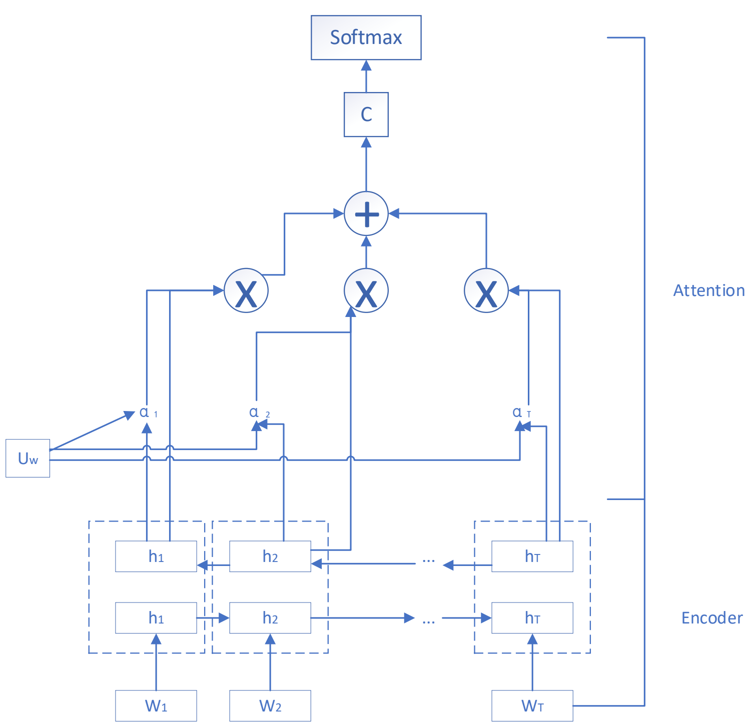

图 2　分类器模型的网络结构

如图 2 所示，分类器模型的网络结构包括 Encoder 和 Attention 两个主体部分，其中 Encoder 部分分为嵌入层（向量序列）、双向 GRU 层；Attention 层的输出经过 Softmax 层后即为分类编码器的分类输出。对于一个关系模板，假设其包含词语序列 $\{w_1,\ldots,w_k,\ldots,w_n\}$，词语对应的词向量为 $\{c_1,\ldots,c_k,\ldots,c_n\}$，词语的位置向量为 $\{v_1,\ldots,v_k,\ldots,v_n\}$，通过向量拼接的方式得到输入数据的嵌入层向量序列 $\{x_1,\ldots,x_k,\ldots,x_n\}$，其中 $x_i=[c_i;\ v_i]$。再把向量序列 $\{x_1,\ldots,x_k,\ldots,x_n\}$ 输入 BiGRU 模块（由前向 GRU 和反向 GRU 构成），来获得隐层状态序列 $\{h_1,\ldots,h_k,\ldots,h_n\}$。即前向 GRU 在第 t 个时刻整合 $\{x_1,\ldots,x_k,\ldots,x_n\}$ 向量序列后，可以生成前向隐向量 $\vec{h}_t$，而反向 GRU 可以生成反向隐向量 $\overleftarrow{h}_t$，这两个隐向量相加则可以得到在第 t 个时刻的隐向量：$h_t=\vec{h}_t+\overleftarrow{h}_t$。

对于分类器模型的训练，本文采用一种常见的 Mean Teacher 半监督学习框架[16]，它通过在标记和未标记样本之间建立一种强大的交互关系，来提升模型的泛化能力。如图 3 所示，在整个分类器模型的训练框架中，关系模板的分类模型主要由学生模型和教师模型组成，它们使用相同的 BiGRU+Attention 深度人工神经网络结构，其中教师模型的参数可以通过学生模型的参数平移得到。高质量样本为人工标注的关系模板及其对应的关系分类标签，可用于学生模型和教师模型的训练；中等质量样本为通过质量相似度得分得到的关系模板及其对应的关系分类伪标签；低质量样本也是通过质量相似度得分得到的关系模板，其伪标签质量较低，因此为无标签数据，但使用了低质量样本的语义信息。三类样本使用不同的损失函数来训练编码器分类模型，其中高质量样本用于训练教师模型（分类器），损失函数采用交叉熵；中等质量样本可同时用于训练学生模型，然后在教师模型指导下获得分类器模型参数，损失函数采用向量的欧式距离；低质量样本可用于训练学生模型，然后在教师模型的指导下获得分类器模型参数，损失函数采用 KL 散度。对高中低三类样本的模型参数进行加权平均，得到一个总的用于关系模板分类的模型，用于预测关系模板对应的关系类型。

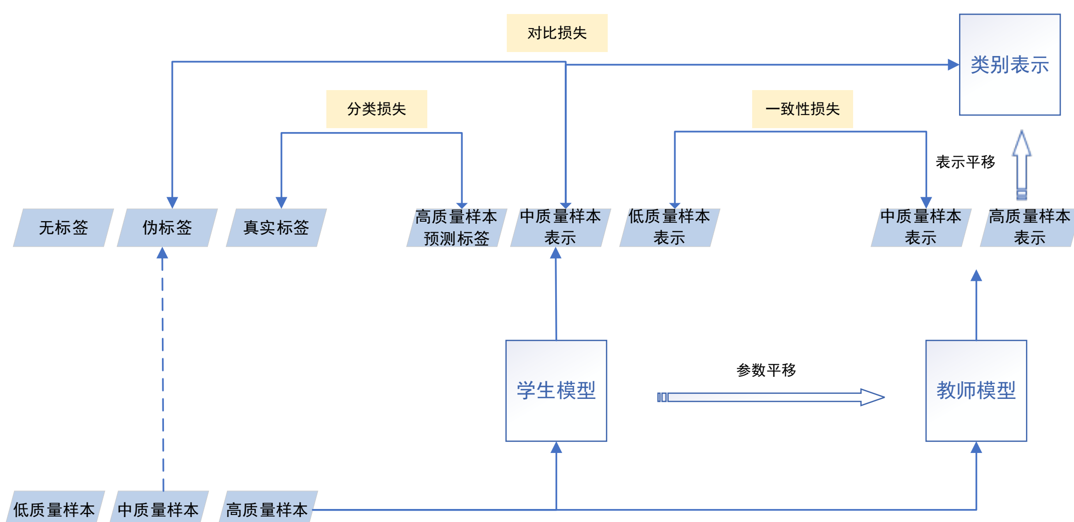

图 3 基于关系模板等级的 Mean Teacher 训练框架

doi:10.3772/j.issn.2095-915x.2024.06.002

4 实验

为了验证基于模板质量划分的半监督 Bi-GRU+Attention 模型在抽取出版产业领域的产业术语和技术术语之间关系的性能和效果，本文以大模型检索的文本结果为基础，总结和设计了产业术语和技术术语之间的关系，并使用基于半监督的深度学习模型对关系模板进行学习和训练，通过实验以验证模型的关系抽取能力。由于技术术语的定义或者百科词条往往不会涉及出版行业，而出版产业术语的定义或者百科词条涉及的具体技术术语也不多，因此要寻找同时出现技术术语和出版产业术语的大量文本极为困难。本文选择大模型搜索平台天工 AI 搜索（https://www.tiangong.cn/）作为工具，使用其“研究”搜索模型，并设计了提问问句，然后手工下载其搜索结果的大模型总结回答文本作为关系抽取的文本语料库。如表 3 所示，对于语料库中同时包含技术术语和产业术语的关系文本，通过依存句法分析可以自动地得到关系模板，同时可以人工标注其关系类型，并且通过产业术语实体的产业链环节得到技术术语所在的产业链的位置（上游、中游还是下游）。

表 3　AI 搜索提问和结果部分文本示例

提问问句	关系文本	关系模板	关系类型	产业链环节
人工智能涉及出版产业的哪些环节？	在印刷和发行环节，人工智能可以用于优化印刷流程，提高印刷效率和质量。	人工智能 用于 优化 印刷 流程 提高 印刷 质量	技术术语实体—应用于—产业术语实体（人工智能—应用于—印刷）	传统出版产业的中游（印刷）
大数据技术在出版产业中有什么具体应用？	出版企业可以利用大数据技术探索新的出版模式，如按需出版、数字出版等，以适应数字化和个性化阅读的趋势。	出版企业 利用 大数据 技术 探索 出版 模式 按需出版	技术术语实体—应用于—产业术语实体（大数据—应用于—按需出版）	数字出版产业的中游（按需出版）
语音识别技术和出版产业有什么关系？	语音识别技术可以用于编辑检索，提高编辑人员的工作效率。	语音识别 技术 用于 编辑 检索 编辑 人员 效率	技术术语实体—应用于—产业术语实体（语音识别—应用于—编辑）	传统出版产业的上游（编辑）

本文设计了 3 类提问问句，使用本文第 2 节得到的部分技术术语和产业术语进行了提问检索，共得到大模型检索的回答文本 1768 篇，从中提取得到同时包含技术术语和产业术语的语句 9591 句。把 9591 句语句文本按照 3:1:1 的比例分为：训练集、验证集和测试集。对每个数据集中的语句文本使用依存句法分析工具（HanLP[24]）自动获取语句中的关系模板，同时根据关系模板的出现频率，手工挑选出少量模板（每一关系类型至少 10 个，至多 50 个）进行关系模板对应的关系类型标注，形成高质量关系模板集合，剩余的关系模板作为未标注关系模板集合。对于未标注的关系模板集合中的关系模板，按照本文 3.1 小节的方法计算关系模板与关系类型（高质量模板）的质量相似度得分，区分为中等质量关系模板和低质量关系模板两类，K 取值为 3。这样得到了高中低质量的三类关系模板，以及这些模板对应的关系类型和词语序列样本，用于后续的半监督深度学习模型的训练和关系抽取。按照本文 3.2 小节的基于关系模板等级的 Mean Teacher 训练框架，对不同等

级质量的关系模板，使用不同的损失函数进行训练和关系抽取。训练集和验证集的语句文本用于模型的学习训练，测试集的语句文本用于进行关系抽取的正确率测试。在模型训练和关系抽取两个阶段，一些超参数的设置有所不同，主要是隐藏层的维度分别设置为 16 和 200，Batch 值分别设置为 16 和 80，关系抽取时中等质量和低质量对应的损失权重分别为 1 和 0.05（训练时权重相同），词向量维度、位置向量维度的设置在训练时和关系抽取时一样，都设置为 200 和 5。

在实验中设置了实体对的关系类型为5类，分别是：应用于、使用、集成、提升、其他。关系抽取的目标就是用训练好的半监督深度学习模型来预测实体间的关系类型，得到知识三元组（头实体，关系，尾实体）。但是实际的数据当中还包含大量未定义的关系类型，为了降低模型的训练难度，因此将未能识别的关系类型标记为其他。此外，部分关系类型（如使用和被使用关系）存在方向性，需要将“应用于”对应的反向关系“使用”加入模型。由于本文使用的语料库及其中的关系标注并没有标准答案和完全人工的标注关系，因此采用基于抽样的人工评估。该方法以抽样的方式从所有三元组中抽取部分样本人工标注，然后计算准确性指标作为关系类型分类的无偏估计。具体来说，按照关系类型分别抽样 200 条知识三元组，然后统计头实体、尾实体、关系类型抽取分类的正确与否。其中，关系抽取的结果需要结合样本进行判断。如果整个三元组（头实体，关系，尾实体），即标注的头实体、尾实体以及关系类型都正确，则认为知识三元组是正确的。通过抽样标注的方法对 5 类关系和模型预测结果进行比对，其准确率结果平均为 66%，如表 4 所示。

表 4　半监督深度学习模型的预测结果准确率

关系类型	样本数	正确数量	错误数量	准确率
应用于	200	136	64	68%
使用	200	153	47	76.5%
集成	200	99	101	49.5%
提升	200	107	93	53.5%
其他	200	165	35	82.5%
平均值	200	132	68	66%

此外，为了验证本文的模型结构是否合理有效，设计了两个消融实验进行验证。一个是在模板质量划分时，仅仅考虑语义相似度而不考虑置信度和语义丰富度，即根据仅依靠最相似的高质量关系模板来确定未标注模板的关系类别，这个消融实验发现置信度和语义丰富度可以带来 1% 的正确率提升。因此，从语义丰富度、置信度等多个角度考虑未标注关系模板的划分，可以提高中等质量模板的健壮性和划分质量，从而提高了整个模型的分类正确率。另一个是消融实验考虑训练集和对比学习对于模型的影响，把中等质量模板集和高质量模板集合起来作为一个训练数据集，并且不考虑不同质量之间的损失函数对比学习。结果发现模型的预测性能在相同的测试集上比原来模型的正确率下降了 1%。这说明虽然直接使用伪标签（中等质量模板的关系类别）来训练分类器也能获得较高的分类正确率，但是加入对比学习后可以减少噪声样本对模型性能的影响，进一步提高模型分类器的关系分类效果。

doi:10.3772/j.issn.2095-915x.2024.06.002

5 结语

本文针对出版产业链，从产业链环节、业务环节、产业术语、技术术语、参与主体、产品服务等六个维度对传统出版产业链、数字出版产业链进行了实体类型划分，并使用无监督的自然语言处理方法，以《编辑与出版学名词》为基准，获得了术语实体 997 个。为了获取产业术语和技术术语之间的关系，本文采用了半监督的深度学习模型方法。首先使用依存句法分析工具从包含术语实体的语句中获得最短依存路径的关键词语，作为关系模板并使用这些词语的词向量表示作为模板的语义表示。然后，利用模板的语义表示和长度等信息计算未标注模板与人工标注模板之间的语义距离及相似度，将关系模板划分为高中低三种类型的模板。借鉴 Mean Teacher 深度学习模型训练框架，使用基于对比学习的损失函数来训练编码器模型。不同质量的关系模板使用不同的损失函数，即高中低三类质量的关系模板分类器参数进行加权平均，得到一个总的分类器模型，可以预测包含实体对的文本语句对应的模板关系类型。通过出版产业的产业术语和技术术语之间的关系抽取实验，发现本文提出的基于模板质量的关系抽取半监督深度学习模型对于 5 种关系类型的预测准确率达到 66%。展望未来的工作，仍需要研究如何通过标注少量的关系模板获得更高准确率的关系类型，或者通过改进深度学习模型框架如加入语义知识等来提高模型的性能。

参考文献

[1] 全国科学技术名词审定委员会 . 编辑与出版学名词 [M]. 北京 : 科学出版社 , 2022: 16, 46, 87.

[2] 薛平 , 李影 , 吴中海 . 基于语言模型增强的中文关系抽取方法 [J]. 中文信息学报 , 2023, 37(7): 32-41.

[3] 任乐 , 张仰森 , 刘帅康 . 基于深度学习的实体关系抽取研究综述 [J]. 北京信息科技大学学报 (自然科学版), 2023(6): 70-79, 87.

[4] 武文雅 , 陈钰枫 , 徐金安 , 等 . 中文实体关系抽取研究综述 [J]. 计算机与现代化 , 2018(8): 21-27, 34.

[5] 郎春雨 , 侯霞 . 基于迁移学习的实体关系抽取技术综述 [J]. 北京信息科技大学学报 (自然科学版), 2022, 37(1): 65-70.

[6] 袁清波 , 杜晓明 , 杨帆 . 限定域关系抽取研究综述 [J]. 计算机系统应用 , 2021, 30(9): 24-40.

[7] 付瑞 , 李剑宇 , 王笳辉 , 等 . 面向领域知识图谱的实体关系联合抽取 [J]. 华东师范大学学报 (自然科学版), 2021(5): 24-36.

[8] 刘彤 , 魏静 , 倪维健 , 等 . 基于半监督学习与 CRF 的应急预案命名实体识别 [J]. 软件导刊 , 2020, 19(3): 35-38.

[9] CHINATSU A, LAUREN H, et al. SRA: Description of the IE^2 system used for MUC-7 [C]// In Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998: 1-14.

[10] FUKUMOTO J, MASUI F, SHIMOHATA M, et al. Description of the Oki system as used for MUC-7 [C]// In Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998: 1-7.

[11] ZELENKO D, AONE C, RICHARDELLA A. Kernel methods for relation extraction[J]. Journal of Machine Learning Research, 2003(3): 1083-1106.

[12] 车万翔 , 刘挺 , 李生 . 实体关系自动抽取 [J]. 中文信息学报 , 2005, 19(2): 2-7.

[13] ZENG D, LIU K, LAI S, et al. Relation classification via convolutional deep neural network [C]// International Conference on Computational Linguistics. The 25th International Conference on Computational Linguistics, 2014: 2335-2344.

[14] SOCHER R, HUVAL B, MANNING C D, et al. Semantic compositionality through recursive matrix-vector spaces [C]// Proceedings of the 2012

Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012: 1201-1211.

[15] 马江微 , 吕学强 , 游新冬 , 等 . 融合 BERT 与关系位置特征的军事领域关系抽取方法 [J]. 数据分析与知识发现 , 2021, 5(8): 1-12.

[16] TARVAINEN A, HARRI V. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results [C]// Advances in Neural Information Processing Systems 30, 2017: 1195-1214.

[17] 郁义鸿 . 产业链类型与产业链效率基准 [J]. 经济与管理研究 , 2005(11): 25-30.

[18] 芮明杰 , 刘明宇 . 产业链整合理论述评 [J]. 工业经济研究 , 2006(3): 60-64.

[19] 魏后凯 . 产业特征、空间竞争与制造业地理集中 [J]. 管理世界 , 2007(4): 68-77.

[20] 郑大庆 , 张赞 , 于俊府 . 产业链整合理论探讨 [J]. 科技进步与对策 , 2011, 28(2): 64-68.

[21] 邓佳佳 . 产业链视角下的数字出版产业发展 [J]. 南昌大学学报（人文社会科学版）, 2014, 45(6): 73-76.

[22] FRANTZI K T, MIMA H, et al. Automatic recognition of multi-word terms: the C-value/NC-value method[J]. International Journal on Digital Libraries, 2000(3): 115-130.

[23] BOUMA G. Normalized (pointwise) mutual information in collocation extraction[J]. Proceedings of GSCL, 2009(30): 31-40.

[24] HE H, CHOI J D. The Stem Cell Hypothesis: Dilemma behind Multi-Task Learning with Transformer Encoders[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 5555-5577.

[25] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Advances in Neural Information Processing Systems 26, 2013: 3111-3119.