基于生成式预训练语言模型的学者画像构建研究*

柳涛[1,2]　丁陈君[1]　姜恩波[1,2]　许睿[1]　陈方[1,2]
（1. 中国科学院成都文献情报中心，成都 610299；2. 中国科学院大学信息资源管理系，北京 100190）

摘要：大数据时代，互联网中以多源异构、非结构化形式存在的学者信息在实体抽取时伴有属性混淆、长实体等问题，严重影响学者画像构建的精准度。与此同时，学者属性实体抽取模型作为学者画像构建过程中的关键模型，在实际应用方面还存在较高的技术门槛，这对学者画像的应用推广造成一定阻碍。为此，在开放资源的基础上，通过引导句建模、自回归生成方式、训练语料微调等构建一种基于生成式预训练语言模型的属性实体抽取框架，并从模型整体效果、实体类别抽取效果、主要影响因素实例分析、样例微调影响分析4个方面对该方法进行验证分析。与对比模型相比，所提出的方法在12类学者属性实体上均达到最优效果，其综合F1值为99.34%，不仅能够较好地识别区分相互混淆的属性实体，对“研究方向”这一典型长属性实体的抽取准确率还提升了6.11%，为学者画像的工程化应用提供了更快捷、有效的方法支撑。

关键词：生成式预训练语言模型；样例微调；学者画像；GPT-3
中图分类号：G351.1　DOI：10.3772/j.issn.1673-2286.2024.03.001
引文格式：柳涛，丁陈君，姜恩波，等. 基于生成式预训练语言模型的学者画像构建研究[J]. 数字图书馆论坛，2024，20（3）：1-11.

学者画像（Scholar Profiling）指对科研工作者的个人信息、科研成果信息、科研社交信息等多维数据进行搜集、整理和展示的过程[1-2]。这个过程旨在创建和维护学者的全面档案，构建精准的学者画像模型，为人才评价、人才引进、人才推荐、跨领域合作等场景提供重要的技术支撑。在学者画像构建过程中，学者信息多以非结构化的文本形式存在，加上学者数量众多，且所属机构及行业各异，其数据来源呈现出明显的非结构化和多源异构特性，为学者画像工作带来不小的困难。与此同时，学者属性实体抽取模型作为学者画像构建过程中的关键模型，目前还存在较高的技术应用门槛，这也对学者画像的应用推广造成了一定的阻碍。以Llama、Baichuan为代表的一批开放性大语言模型的面世与应用，使得便捷、快速、高效地构建学者画像成为一种可能。为此，本文尝试将学者画像构建与生成式预训练语言模型相结合，重点考虑模型抽取精度的提升以及应用门槛的降低，以期为学者画像的工程化应用提供更快捷、有效的方法支撑。

1　研究综述

学者画像构建方法一直是学者画像领域的研究重点，通常建立在真实学者数据之上，运用数据挖掘、信息抽取等技术刻画不同学者的特征属性[2]，完成学者属性内容的结构化处理。从流程上来看，学者画像构建主要包含源数据确定与采集、学者画像维度设计[3]、多源数据融合[4]、学者属性自动抽取[5]以及学者画像可视化展示[6]等阶段。从现有研究来看，构建高质量学者画像的关键在于学者特征属性的自动准确抽取，即学者属性实体抽取模型的精准度。依据方法模型的不同，学者

收稿日期：2023-12-12
*本研究得到“西部之光”人才培养计划“基于模式创新的医药生物产业科技服务体系研发及应用示范”（编号：E1C0000401）、中国科学院成都文献情报中心创新基金项目“生物-信息科技情报领域智慧数据体系建设”（编号：E1Z0000101）资助。

画像构建中的属性实体抽取模型可分为基于规则或词典、基于统计机器学习和基于深度学习3类。

基于规则或词典的方法构建模式规则或者属性词典，利用规则或词典从原始文本中抽取属性信息。池雪花[7]基于触发词、正则表达式等方法制定学者属性抽取规则，完成学者性别、邮箱、职位等属性的抽取。这类方法中词典和规则的完备性是决定抽取效果的关键影响因素。大数据环境下，面对多源异构的学者个人信息，人工方式难以覆盖所有的属性词典以及抽取规则，因此，这类方法通常应用于文本结构较为统一的学者数据抽取。

基于统计机器学习的实体抽取模型通常将学者属性实体抽取任务视为标签分类问题或者序列标注问题。这类方法在可移植性和泛化能力上有着良好的表现。Tang等[8]提出一种基于树结构的条件随机场（Conditional Random Field，CRF）模型，从学者文档中抽取基础信息，抽取效果的F1值为86.7%。Gu等[9]提出用于学者邮箱和性别抽取的MagicFG模型，并在AMiner平台提供的2 000名学者数据上取得了较好的抽取结果。范晓玉等[10]结合哈尔滨工业大学语言技术平台（Language Technology Platform，LTP）的命名实体处理工具和学者属性词典，对学者的姓名、籍贯、机构、出生日期等属性进行抽取。王锐杰[11]采用条件随机场模型对学者学术简历中的机构、头衔、荣誉进行标签标注，平均准确率为74.2%，并指出长实体的存在进一步增加了学者实体抽取的难度。以上方法模型虽然取得了一定程度上的性能提升，但是模型效果与输入数据特征之间存在较高关联度。构建合适的数据特征来表征学者文本语义是其关键，而这一过程通常较为繁杂，同时难以囊括所有语义特征，从而导致基于统计机器学习的模型在文本的语义捕获、上下文建模等方面存在不足。

基于深度学习的方法使多级自动特征表征学习成为可能。研究人员开始从语义表征、信息增强等方面优化原有方法模型。张亚楠等[12]先结合主题模型与长短期神经网络（Long Short Term Memory，LSTM）分别捕获科研行为数据中的局部和全局信息，再采用TSP（Two Stage Persona）框架确定科研学者画像标签。李微[13]基于DOM（Document Object Model）树完成学者主页文本块的切割抽取，借助文本块的实体特征和语义特征完成文本分类，最终从相应类别的文本块中提取学者画像相关属性。BERT（Bidirectional Encoder Representation from Transformers）模型的出现使得语义表征增强的建模道路得到进一步的延伸[14]。滕倚昊[15]引入BERT完成文本语义的高质量表征，结合双向长短期神经网络（BiLSTM）的上下文建模能力，提出BERT-BiLSTM-CRF模型进行学者属性实体的抽取，实验结果表明该方法模型表现出了超越传统模型的性能，优化了学者画像构建过程中关键任务的模型表现，具有较高的工程应用价值。

ChatGPT颠覆性的应用效果以及开放统一的模型架构让其背后的支撑模型——生成式预训练语言模型（Generative Pre-Trained Transformer，GPT）迅速引起了研究人员的关注[16-18]。这种模型采用两段式训练模式——预训练（Pre-Training）和微调（Fine-Tuning）：模型在具备通识能力之后，通过特定领域的数据微调获得一定的专业能力。与语义表征增强的建模道路不同，生成式预训练语言模型有着显著的训练数据、模型规模等方面的优势，能从大规模自由文本中进一步捕获和学习其中的语言规律、词汇特征等文本信息，从而具备更强大的文本理解与生成能力。此外，现今大部分大语言模型均以开源形式出现，这种形式为这类强大基座模型的应用研究提供了重要基础。

从上述研究与实践来看，学者属性实体的抽取方法一直在发展进步，其研究重点多为属性实体抽取结果的逐步优化，但笔者也注意到学者属性实体抽取目前仍然存在两个缺陷：①面对长实体、属性混淆的复杂内容时，模型无法充分理解文本语义，难以取得理想的抽取效果；②学者属性实体抽取模型的精度提升基于不同模型结构的设计与结合，在模型开放程度较低的情况下，复杂的模型结构也提高了模型的应用门槛。本文基于上述两个缺陷，拟引入开放性较高且具备较高知识水平的生成式预训练语言模型以降低模型应用的技术门槛，同时结合引导句建模、自回归生成方式和训练语料微调的方法，聚焦长实体识别与属性混淆两方面的问题，进一步提升学者属性实体抽取的精准度。

2 研究设计

随机选择国内高校发布的学者个人中文主页作为语料数据来源。在语料准备阶段主要完成学者名单和主页确定、主页文本采集和预处理、目标属性标注、标注格式转换以及语料划分等流程，为后续阶段研究提供数据支撑。

实验环境是OpenAI公司的Playground。OpenAI Playground允许用户以不同的方式与GPT-3交互，并查看模型如何真实地生成响应。同时，Playground也是一个在线工具，用户无需安装即可通过互联网连接访问，由此保证了研究与实验结果的可扩展性以及后续工程化推广的便捷性。

2.1　基本概念

长实体识别是指在学者画像过程中对一些以自然语言描述的、以句子为长度单元的文本内容进行实体抽取。长实体识别的困难在于实体组成灵活多变、实体边界模糊。学者研究领域、研究兴趣、所获奖项等内容是较为典型的长实体类型，例如以下段落中的粗体部分。

“姓名：×××。工作单位：四川大学机械工程学院。职称：副研究员。导师情况：硕士生导师。邮箱：chaolangchen@scu.edu.cn。招生方向：机械工程。教育背景及工作经历。教育背景。2016.09—2021.07清华大学，机械工程系，博士。2019.10—2020.04伦敦大学学院，化学系，研究助理。2012.09—2016.07东北大学，机械工程及自动化，本科。工作经历。2021.07至今四川大学，机械工程学院，副研究员/副教授。总体介绍。主要从事**机械表界面技术、仿生微纳制造、油水分离技术、多相流模拟等方面的研究**。作为负责人主持国家自然科学基金青年项目、四川省重点研发、校地合作研发和国家重点实验开放基金等多个科研项目；作为学术骨干参与国家重点研发、国家自然科学基金面上项目等多项国家级项目；获中国机械工业科学技术发明二等奖（2021年）。”

属性混淆是指学者属性中存在多个与目标属性词相混淆的其他实例词，同时目标属性词可能在文本不同位置多次出现。最为典型的属性混淆现象在学者的职称、工作单位以及毕业院校和时间中有所体现。以各阶段的毕业院校为例，学者简历中可能出现多个表述院校的属性实体，例如以下段落中的粗体部分。

“×××。**四川大学**特聘研究员/教授、博士生导师。2015年国家自然科学基金优秀青年科学基金项目获得者。1999年毕业于**北京大学**生命科学学院获学士学位，2009年毕业于（美国）**密歇根大学**细胞与发育生物学系，获博士学位。2010—2014年在（美国）冷泉港实验室、纪念斯隆凯特琳癌症中心的**Scott Lowe实验室**从事博士后工作。2014年8月受聘于**四川大学**。”

对于模型来说，其难点在于既要准确完整地抽取出表示院校的实体片段，还要辨析各个院校实体所属的实体类型，这对模型的语义辨析能力有较高的要求。

2.2　模型样例微调

模型样例微调阶段一方面完成生成式实体抽取标注方式的转换，另一方面促使模型完成学者领域知识与属性实体抽取规范的学习。具体模型架构如图1所示，其中：y、g、x分别表示属性实体文本、引导句、待抽取文本中的字符；k、r、m分别表示属性实体文本、引导句、待抽取文本中的字符数量；Trm表示Transformer单元。生成式预训练语言模型有着较为统一、完善的模型

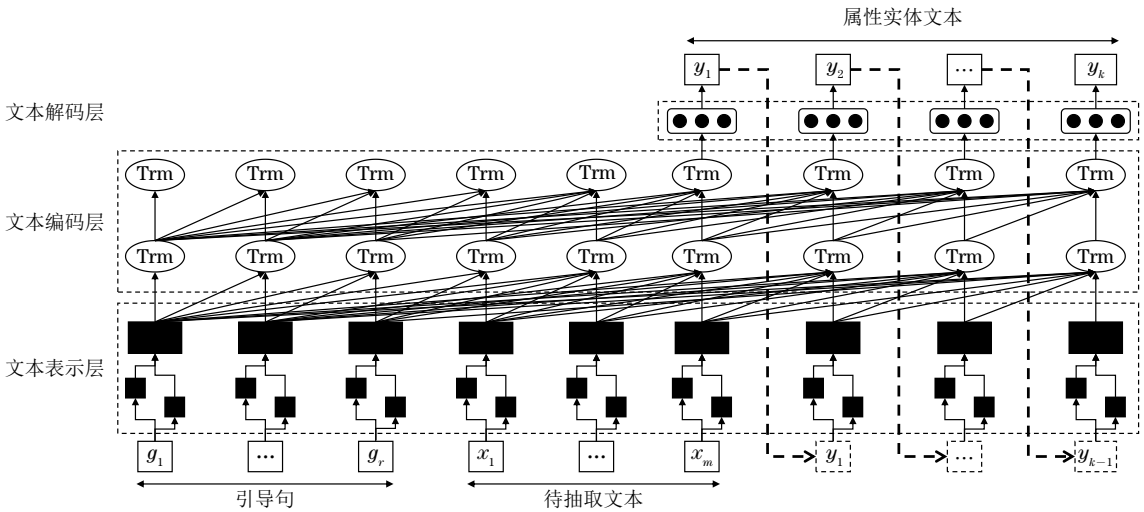

图1　生成式学者属性实体抽取模型框架

基础结构，本研究在此基础上融入引导句建模机制，形成生成式学者属性实体抽取模型框架。该模型框架针对属性混淆问题，为每一学者属性增加相应的引导句，并利用Transformer单元含有的注意力机制，捕获引导句、学者描述文本、目标属性实体的语义联系，增强模型在特定引导句下对学者目标属性的辨析能力，引导模型定位学者描述文本中的正确属性实体。

一方面，与序列标注模型不同，生成式属性实体抽取模型精简了原有文本序列到标签序列以及标签序列到文本序列的转换流程，减少了学者属性实体抽取框架中的中间流程，为实现真正意义上的端到端属性实体抽取奠定基础。另一方面，生成式属性实体抽取模型可以在属性混淆情况下有效避免目标属性的重复抽取，而序列标注模型可能会将学者描述文本中的所有目标属性实体都标注出来，从而增加额外工作。

针对属性混淆问题，具体计算流程如下。

文本表示层的计算流程如式（1）所示。

$$\boldsymbol{h}_0 = \boldsymbol{U}\boldsymbol{W}_{\mathrm{e}} + \boldsymbol{W}_{\mathrm{p}} \tag{1}$$

式中：$\boldsymbol{U} = (G, X)$ 为引导句G和待抽取文本X拼接后的向量表示；引导句 $G = (g_1, g_2, \cdots, g_r)$，待抽取文本 $X = (x_1, x_2, \cdots, x_m)$；$\boldsymbol{W}_{\mathrm{e}}$为语义嵌入矩阵；$\boldsymbol{W}_{\mathrm{p}}$为位置嵌入矩阵；$\boldsymbol{h}_0$为文本表示层的计算结果。

基于堆叠Transformer单元的文本编码层的计算流程如式（2）所示。

$$\boldsymbol{h}_i = \mathrm{transformer\_item}(\boldsymbol{h}_{i-1}) \forall i \in [1, n] \tag{2}$$

式中：transformer_item表示Transformer单元中的Decoder计算结构；i表示堆叠的Transformer层数；$\boldsymbol{h}_i$表示第i层Transformer的计算结果。

文本解码层的计算流程如式（3）所示。

$$P(y \mid g_1, \cdots, g_r, x_1, \cdots, x_m) = \mathrm{softmax}(\boldsymbol{h}_n^{(r+m)} \boldsymbol{W}_{\mathrm{y}}) \tag{3}$$

式中：$\boldsymbol{W}_{\mathrm{y}}$为文本解码层的参数矩阵；$\boldsymbol{h}_n^{(r+m)}$ 为第$r+m$位置经过n层Tansformer计算后的隐变量；P为条件概率分布函数。

针对长实体问题引入自回归生成方式。自回归的生成式将已生成的属性文本逐一加入原文本，结合生成的内容和原文内容来生成后续学者属性内容。由于长实体本身的字符数量较多，整体上下文连贯性较强，这种自回归的生成方式为长实体边界的确定提供额外的上文语义信息，为准确完整生成长实体提供了理论基础。基于自回归生成方式的计算公式如式（4）所示。

$$P(Y \mid X) = \prod_{t=1}^{k} P(Y_t \mid Y_{1:t-1}, X) \tag{4}$$

式中：$Y = (y_1, y_2, \cdots, y_k)$，$Y_{1:t-1}$ 表示属性实体文本中第1到第$t-1$位置的所有文本。

为了提升生成式预训练语言模型在学者领域文本中的属性实体抽取表现，采用样例微调的方式，通过在特定数据集和自然语言处理任务上重新调整模型参数，使模型适应新的知识领域以及任务规范。

模型参数的调整与训练过程中的训练目标直接相关。为促使生成式预训练语言模型充分理解、学习学者领域知识以及属性实体抽取规范，以最大化学者目标属性实体标准序列的生成概率为训练目标。通过学者领域的属性实体标注样例来对原有生成式预训练语言模型进行样例微调，不断修正生成式预训练语言模型的参数集φ，完成学者领域的增量式学习。训练目标的具体数据形式化表示如式（5）所示。

$$\varphi = \mathrm{argmax}\left[\sum_{(x,y)} \log P(y \mid x_1, \cdots, x_m; \varphi)\right] \tag{5}$$

2.3 实体抽取

经过模型样例微调阶段之后，针对各类学者属性都已经形成对应的实体抽取微调模型。因此，在实体抽取阶段，分别将测试语料中的学者原始文本序列逐一输入已经微调完毕的各个学者属性实体抽取模型，完成对各类学者属性实体的抽取操作。具体计算公式与式（1）~（3）一致。

2.4 结果评估阶段

采用实体抽取领域常用的准确率（Precision）、召回率（Recall）和F1值3个评测指标来进行模型评估，具体计算公式如式（6）~（8）所示。

$$R_{\mathrm{Precision}} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}} \tag{6}$$

$$R_{\mathrm{Recall}} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}} \tag{7}$$

$$S_{\mathrm{F1}} = \frac{2 \times R_{\mathrm{Precision}} \times R_{\mathrm{Recall}}}{R_{\mathrm{Precision}} + R_{\mathrm{Recall}}} \tag{8}$$

式中：$R_{\mathrm{Precision}}$、R_{Recall}、S_{F1}分别代表准确率、召回率、F1值；N代表数据数量；TP为真正例，代表真实的数据是正例，并且预测值也是正例；FN为假反例，代表

真实的数据是正例，而预测值为反例；FP为假正例，代表真实的数据是反例，而预测值是正例。

根据对比实验原则，对学者属性实体抽取结果采用自动标注的方式，将抽取结果转换为标签序列，从而保证结果形式与现有实体抽取模型一致。对于抽取错误的实体结果，采用对应的属性实体标签替换原有文本语料中的非属性实体字符标签，还原模型标注错误。在此基础上，分别从准确率、召回率以及F1值3个指标出发对模型进行综合评估，以验证生成式预训练语言模型在学者属性实体抽取上的可行性与有效性。

3　实证研究

3.1　实验语料库的构建及分析

3.1.1　实验语料库的构建

随机选择国内多所高校中文官网发布的学者个人主页作为主要语料数据来源，一方面保证了学者信息的真实性与可靠性，另一方面保留了不同学者的个人主页在页面结构与内容上的差异，从而使实验语料库具备一定的现实代表性。选择云南大学、浙江大学、四川大学、北京大学、江西农业大学等15所高校经济学、工学、管理学、农学、理学、法学等多个学科领域的1 000位学者的个人主页作为实验语料，确保实验语料具备多源异构性。利用Python编程的selenium库，依据学者主页表进行主页文本采集，并通过正则表达式、人工审核等方式对其中特殊字符、格式等进行清除整合，完成文本预处理操作，得到学者原始语料文档。

在学者属性标注方面，结合现有学者画像研究调研以及实际需求，确定个人主页中的学者基础信息属性为邮箱、职称、最高学历、导师资格、工作单位、研究方向、教育背景。其中，教育背景作为一项综合性的属性实体项，通常分为学士、硕士、博士3个阶段的经历，针对每一阶段，又有毕业时间和毕业院校两项重要信息，为此将教育背景细分为学士时间、学士院校、硕士时间、硕士院校、博士时间、博士院校6种实体类型。最终学者基础信息属性共计12项。在此基础上制定相关的学者属性实体标注规范，由2名情报学硕士研究生在学者属性实体标注规范基础上，各标注50位学者数据，并相互核验标注结果，校正标注规范。重复实验标注步骤后使两人的标注结果接近一致，完成剩余学者属性信息标注，形成标注语料文档。对于有多个正确属性词的学者属性，标注第一个出现的即可。对于没有正确属性词的学者属性，采用no-answer进行标注。具体学者基础信息属性类型及样例如表1所示，其中粗体部分为该实体类别下待抽取的属性文本片段。

表1　实验语料的实体类别和样例

实体类别	子类	样例
邮箱	无	电子邮件：**ffang@pku.edu.cn**。个人简历。麦戈文脑科学研究所常务副所长
职称	无	**教授**，博士生导师，四川省杰青
最高学历	无	**博士**、教授、博士生导师
导师资格	无	药理学，**硕士生导师**
工作单位	无	工作地址：**复旦大学**药学院
研究方向	无	主要从事**机器视觉检测和微纳测试技术方面的研究**
教育背景	学士时间	**1983.07—1987.07**武汉大学化学系，学士
	学士院校	1983.07—1987.07**武汉大学**化学系，学士
	硕士时间	**1987.07—1990.07**中国科学院感光化学研究所，硕士
	硕士院校	1987.07—1990.07**中国科学院感光化学研究所**，硕士
	博士时间	**1994.07—1997.07**中国科学院感光化学研究所，博士
	博士院校	1994.07—1997.07**中国科学院感光化学研究所**，博士

在标注格式转换以及语料划分方面，与原有序列标注方式不同，采用文本生成的方式完成学者属性实体抽取，采用Prompt+Completion的问答形式作为标注数据格式。为此，设计了文本生成范式下的标注语料转换模板，将人工标注完毕的学者语料填充到语料转换模板中，完成标注形式的转换（见图2）。对于不同的学者属性，同样采用模板填充的方式形成不同学者属性下的模型引导句（见表2）。最终，按照7∶3的比例将转换后的标注语料划分为训练语料与测试语料，完成生成范式下的训练语料构建。

3.1.2　实验语料库的代表性验证

从学者属性实体的长度分布情况来看，大部分学者属性实体的字符数量在5个以上。其中，研究方向属性文本字符占比较大，平均实体长度为22.84个字符，最长为93个字符，存在较为明显的长实体现象（见图3）。

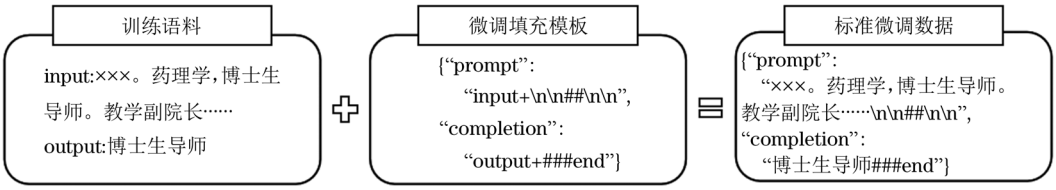

图2　标准语料构建流程

表2　学者属性实体对应的模型引导句

属性实体类型	模型引导句
邮箱	从上述文本中抽取出表示邮箱的文本片段
职称	从上述文本中抽取出表示职称的文本片段
最高学历	从上述文本中抽取出表示最高学历的文本片段
导师资格	从上述文本中抽取出表示导师资格的文本片段
工作单位	从上述文本中抽取出表示工作单位的文本片段
研究方向	从上述文本中抽取出表示研究方向的文本片段
学士时间	从上述文本中抽取出表示学士时间的文本片段
学士院校	从上述文本中抽取出表示学士院校的文本片段
硕士时间	从上述文本中抽取出表示硕士时间的文本片段
硕士院校	从上述文本中抽取出表示硕士院校的文本片段
博士时间	从上述文本中抽取出表示博士时间的文本片段
博士院校	从上述文本中抽取出表示博士院校的文本片段

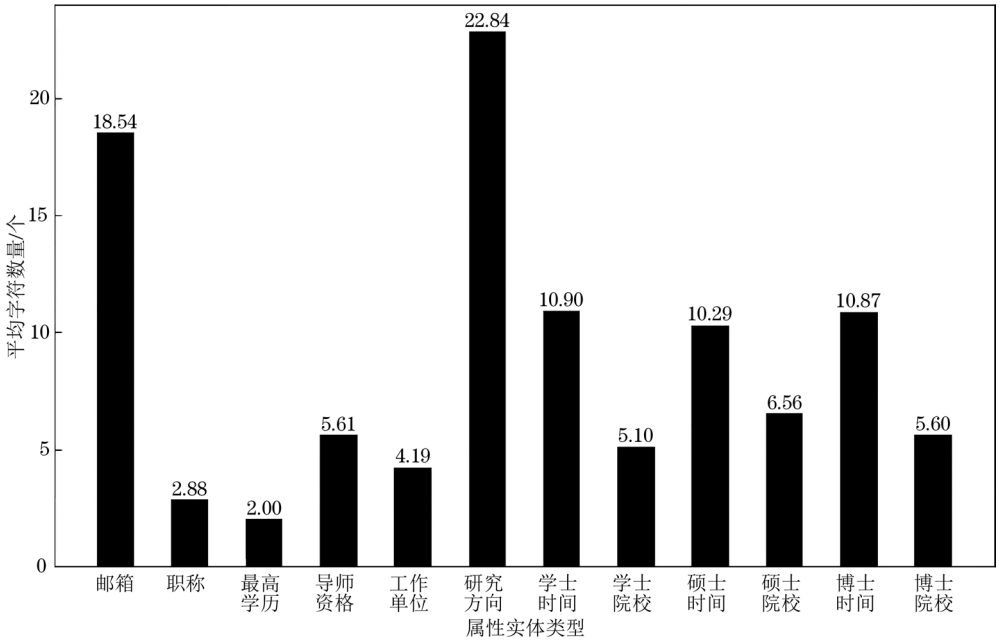

图3　命名实体长度分布

3.2　实验参数与实验环境

依据提出的生成式学者属性实体抽取模型框架，并以GPT-3 Babbage模型为基础，完成相关实验。GPT-3 Babbage模型包含很多参数，其中主要的参数为Max_tokens、Top p、Temperature、Frequency_penalty、Presence_penalty，参数设置见表3。其中Max_tokens表示文本生成时的最大字符数量，目前GPT-3的Max_tokens最大为2 048。分析实验语料中最长实体长度为93个字符，为保证模型具备一定的扩

展性，实验设置256作为Max_tokens值。Top p控制生成文本的概率分布范围，通过排除低概率的词语来选择更合理的生成字符，取值范围为0~1，取值越高生成的内容文本越多样化。Temperature用于控制生成文本的随机性，值越高，生成的文本字符越随机。考虑到语料中不同学者的研究方向实体之间具备较高的差异性，为保证生成内容的多样性和随机性，将Top p和Temperature分别设置为1和0.7。Frequency_penalty主要用于词频惩罚，值越高，生成文本含有重复词语的概率越小，取值范围为0~10。学者属性实体中可能存在较多重复词，因此不对其进行惩罚，其Frequency_penalty设置为0。Presence_penalty主要用于存在惩罚，取值范围为0~10，值越低，则越容易生成前文中存在的特定词语。本研究主要抽取前文中已有的实体文本内容，因此将Presence_penalty设置为0。

表3　实验参数设置

Max_tokens	Top p	Temperature	Frequency_penalty	Presence_penalty
256	1	0.7	0	0

3.3　实验结果与分析

选取隐马尔科夫模型（Hidden Markov Model，HMM）、CRF和BERT-LSTM-CRF模型作为对比基线模型。HMM作为命名实体识别任务中的经典模型之一，训练速度快，复杂度低，但容易在训练过程中陷入局部最优解[19]。CRF模型考虑序列标注问题的全局最优解，利用上下文的语义特征完成标签实体的分类，能有效避免标签偏置问题[20]。BERT-LSTM-CRF模型主要采用文本表示层、序列编码层和标签解码层3层网络结构完成命名实体建模，是目前命名实体识别领域效果较优的模型之一[21]。

3.3.1　模型整体效果

按提出的评估指标来测试4个模型在准确率、召回率和F1值方面的表现。测试结果如表4所示。其中模型1作为早期经典模型，其F1值为73.95%，具有一定的识别效果。模型2在模型1基础上考虑到标签序列的全局最优，其识别效果的F1值为89.82%，相较于模型1提升15.87%，提升效果明显。模型3为主流的深度学习模型，相较于前两个模型在整体网络结构、模型设计上都更为复杂、完善，其F1值达到了96.68%，这说明深度学习模型在学者属性实体抽取方面精准度更高。模型4的F1值则在模型1、2、3的基础上分别提升了25.39%、9.52%、2.66%，说明本文提出的模型框架相较于以往模型，在总体效果上表现更优。

从建模形式来看，模型1、2、3采用常见的序列标注形式，而模型4采用文本生成方式完成学者属性实体抽取，以类似“阅读理解—任务回答”的方式在学者主页文本内容基础上直接生成学者属性实体。这种方式精简了从原有文本序列到标签标注序列再到学者属性实体文本序列的转换流程，完成原文文本到学者属性实体文本的直接转换。

从模型绝对性能来看，模型4的F1值为99.34%，突破了原有深度学习模型的抽取上限。其在模型3的高准确率基础上，还能有2.66%的效果提升，达到基本准确地抽取出学者主页中含有的各属性实体的目标，为学者画像构建提供扎实的技术基础。

3.3.2　实体类别抽取效果

分别对12种学者属性实体抽取的F1值进行计算分析，结果如图4所示。在邮箱、工作单位、学士时间、学士院校、硕士时间、硕士院校、博士时间、博士院校等8项学者属性抽取方面，提出的生成式学者属性实

表4　模型的整体结果评估

单位：%

代号	实验模型	准确率	召回率	F1值
1	HMM	84.92	69.60	73.95
2	CRF	89.55	90.45	89.82
3	BERT-LSTM-CRF	95.87	97.51	96.68
4	生成式学者属性实体抽取模型	99.34	99.33	99.34

体抽取模型与BERT-LSTM-CRF模型的抽取效果类似。而在属性混淆和长实体现象表现较为突出的属性实体上，例如职称、最高学历、导师资格、研究方向等，所提模型的抽取效果则优于BERT-LSTM-CRF模型。实验结果表明：一方面，采用的引导句加样例微调的组合建模形式有效增强了模型在特定引导句下对学者目标属性的辨析能力；另一方面，自回归的生成方式能联合已生成的实体文本和原文本来迭代生成后续的学者属性实体，有效解决了长实体的抽取难题。

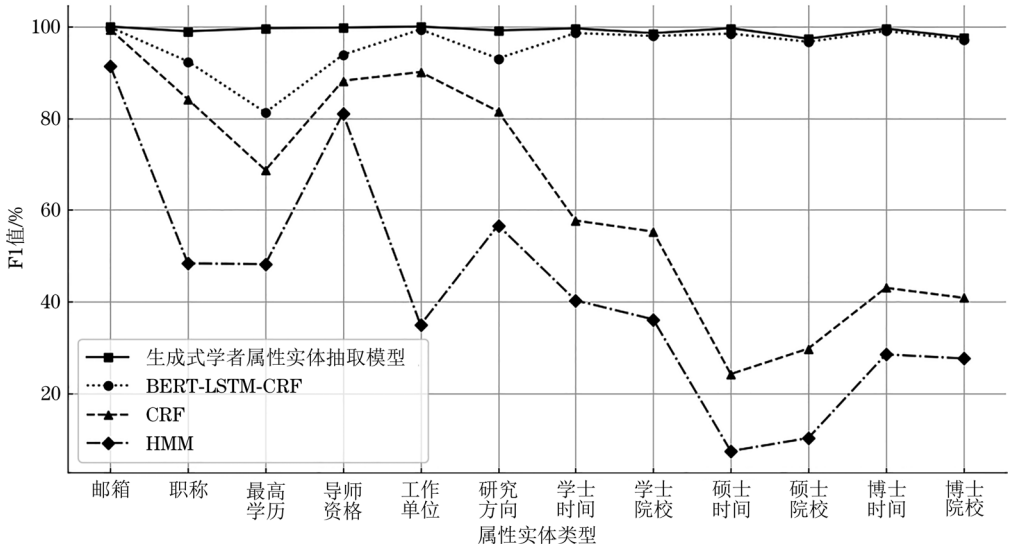

图4　分类分析结果

3.3.3　主要影响因素实例分析

对学者领域中存在的属性混淆与长实体难题进行具体的抽取实例分析，分别对两类问题的具体表现、解决方式以及实验结果进行阐述。

从属性混淆现象来看，职称和学历方面，学者主页文本中存在多个同类型的实例词，导致明显的属性混淆现象。例如：学者个人主页可能描述学者的工作经历，其中包含“教授”“副教授”“讲师”等同属于职称的实例词；学历一般是指学者的最高学历，而在个人主页中可能会出现“学士”“硕士”“博士”等实例词。针对这种属性混淆现象，在具备通用知识的生成式预训练语言模型的基础上，以最大化目标学者属性实体标准序列的生成概率为训练目标，同时结合属性引导句，驱动语言模型进一步学习学者领域的属性实体标注语料样例，定向强化模型的学者文本领域知识。从图5和图6可以看出，提出的方法模型分别抽取出了正确的“教授”职称和“博士”学历，有效区分同类型的实例词“研究员”或“学士”“硕士”等，提升了学者领域文本语义理解和辨析的有效性。

从长实体现象来看，研究方向实体类型不仅整体字符数偏多，还存在多样化特征，同一学者可能使用有多个语义不相关的句子来表示其研究方向，因此，识别研究方向实体边界存在较大难度。采用自回归生成的实体抽取方式可在原有最优模型的基础上将准确率提升6.11%，提升效果显著。如图7所示，BERT-LSTM-CRF、CRF模型在HMM模型基础上对实体边界的判别准确率略有提升，但“本构关系”与前文的语义差异过大，仅所提的方法模型可完整、准确地抽取出实例中存在的长实体。

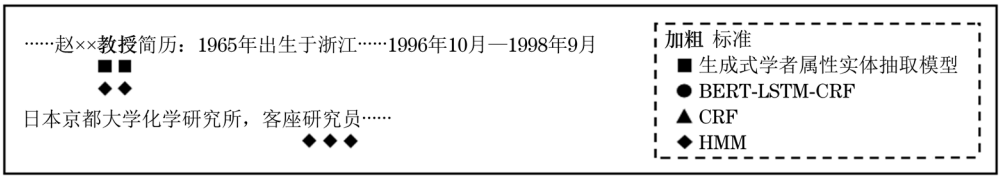

图5　职称实例分析

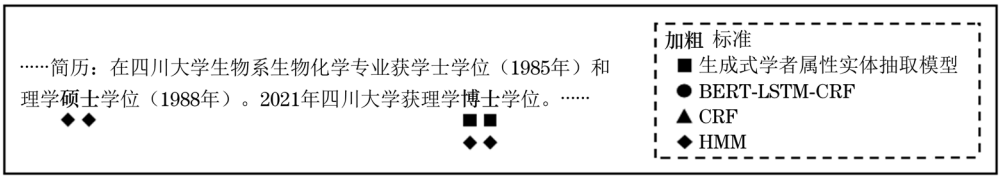

图6　学历实例分析

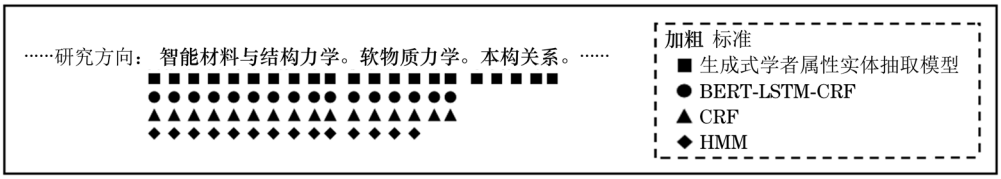

图7　研究方向实例分析

3.3.4　样例微调影响分析

生成式预训练语言模型具备较高的文本对话能力，通过这种文本对话的方式可以进一步分析基础模型在样例微调前后的能力变化。以研究方向、导师资格、职称、最高学历等4项学者属性为例，在OpenAI开放的Playground平台，分别选择原始Babbage模型以及样例微调后的Babbage模型对学者描述文档进行属性抽取实验，文档部分内容如下。

“×××：药物化学博士，教授，博士生导师；北京大学药学院化学生物学系主任。电话：（010）-82801714（办公室）。传真：（010）-82805496。E-mail：zjli@bjmu.edu.cn。通讯地址：北京市海淀区学院路38号北京大学医学部药学楼633。邮编：100191。1982—1987年就读于北京医科大学药学院化学专业获理学学士学位，1987—1992年就读于北京医科大学药学院药物化学专业获药物化学博士学位，后留校任教从事教学和科研工作。1999年被聘为教授，2001年被聘为博士生导师。长期以来从事糖化学、糖化学生物学及相关创新药物的研究。先后获得了国家自然科学基金重大研究计划、重点课题、面上项目、科技部‘973’、科技部创新药物重大专项、国家新药研究基金、卫生部科研基金、教育部创新团队等40余项基金的支持。已发表学术论文150余篇和专著1部，获批中国发明专利及国际专利20余项。担任中国药学会药物化学专业委员会委员，北京药学会药物化学专业委员会主任委员；为《Chinese Chemical Letter》《中国药物化学杂志》《中国药学》（英）等杂志编委。已指导研究生80余人获得学位。曾获国家教委科技进步三等奖及中国药学会Servier青年药物化学奖。目前承担的课程有‘有机合成’（药学院本科）、‘糖化学’（药学院研究生）。先后获得北京大学‘十佳教师’‘优秀教师’及医学部‘十佳教师’、研究生‘良师益友’‘我身边的好老师’等称号。”

结果显示，原始Babbage模型在抽取形式和抽取内容上与样例微调后的模型存在明显差异。以研究方向、导师资格、最高学历3项为例，对比模型微调前后的抽取效果，如表5所示。

在形式上，原始Babbage模型的生成内容不符合实体抽取任务规范，模型不能直接给出具体的学者属性实体，而是给出一段自由交流文本，故不能有效满足实体抽取的结构化需求。而样例微调后的Babbage模型准确地给出了学者属性实体并完成直接抽取任务，结果形式更加符合实体抽取任务规范。

表5　样例微调前后抽取结果对比

属性实体类别	微调前抽取结果	微调后抽取结果
研究方向	1.中国药学会药物化学专业委员会学术委员会秘书职责1.1学术委员会管理委员会秘书职责1.2学术委员会担任1.3学术委员会工作机构	糖化学、糖化学生物学及相关创新药物的研究
导师资格	作者：张艳霞，张秀萍、凤福永，等，韩素泉酸对多义单核苷酸对二乙基苷酸的水稻表达机制：糖果有机化	博士生导师
最高学历	教育和研究	博士

在内容上，原始Babbage模型所生成的文本内容与学者领域的相关性较小，模型生成的文本更倾向于通用文本内容，表现出较低的学者领域知识水平。样例微调后的Babbage模型能有效理解给出的学者主页全文，在此基础上针对属性实体的抽取需求，实现领域实体的准确抽取。从生成内容对比来看，样例微调有效推动了预训练语言模型对学者领域知识的增量学习。

4 结语

本文以现有开放性较强、基础知识水平较高的生成式预训练语言模型为基础，探索了生成式预训练语言模型在学者画像领域的具体应用。结果表明：一方面，融合引导句建模、自回归生成方式以及样例微调的模型框架不仅继承了现有大规模预训练语言模型的基础能力以及规模优势，还实现了对原有模型学者领域知识的补充以及属性实体抽取的优化，进一步解决了长实体、属性混淆所带来的抽取难点，取得了F1值为99.34%的良好效果；另一方面，研究所采用的基础模型与实验环境均为开放获取资源，具有良好的通用性和可移植性，降低了学者画像工程化应用的技术门槛，为推广学者画像的相关研究提供了更有效的方法支撑。综合来看，这类生成式预训练语言模型的应用研究将为包括但不局限于学者画像领域的各行各业提供更便捷、有效的实体抽取解决方案，进一步推动学者画像领域的模型变革以及上层应用。

参考文献

[1] 王世奇，刘智锋，王继民. 学者画像研究综述[J]. 图书情报工作，2022，66（20）：73-81.

[2] 袁莎，唐杰，顾晓韬. 开放互联网中的学者画像技术综述[J]. 计算机研究与发展，2018，55（9）：1903-1919.

[3] 董文慧，熊回香，杜瑾，等. 基于学者画像的科研合作者推荐研究[J]. 数据分析与知识发现，2022，6（10）：20-34.

[4] 范晓玉. 基于多源科技管理数据的重大项目团队成员推荐研究[D]. 西安：西安电子科技大学，2018.

[5] GENG Q，CHUAI Z A，JIN J. Automatic construction of academic profile：a case of information science domain[J]. Journal of Information Science，2023，49（1）：207-232.

[6] 陈玲洪，潘晓华. 基于知识图谱和读者画像的图书推荐研究[J]. 数据分析与知识发现，2023，7（12）：164-171.

[7] 池雪花. 学者精准画像的自动构建研究[D]. 南京：南京理工大学，2019.

[8] TANG J，YAO L M，ZHANG D，et al. A combination approach to web user profiling[J]. ACM Transactions on Knowledge Discovery from Data，2010，5（1）：2.

[9] GU X T，YANG H，TANG J，et al. Web user profiling using data redundancy[C]//2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining（ASONAM）. 2016：358-365.

[10] 范晓玉，窦永香，赵捧未，等. 融合多源数据的科研人员画像构建方法研究[J]. 图书情报工作，2018，62（15）：31-40.

[11] 王锐杰. 基于多源信息融合的科研学者画像及应用研究[D]. 成都：电子科技大学，2020.

[12] 张亚楠，黄晶丽，王刚. 考虑全局和局部信息的科研人员科研行为立体精准画像构建方法[J]. 情报学报，2019，38（10）：1012-1021.

[13] 李微. 一种用于描述学者画像的信息抽取系统[D]. 南京：东南大学，2020.

[14] DEVLIN J，CHANG M W，LEE K，et al. BERT：pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2023-11-12]. http://arxiv.org/abs/1810.04805v2.

[15] 滕倚昊. 基于神经网络的学者画像研究与应用[D]. 南京：东南大学，2020.

[16] 张智雄，于改红，刘熠，等. ChatGPT对文献情报工作的影响[J]. 数据分析与知识发现，2023，7（3）：36-42.

[17] 张智雄，曾建勋，夏翠娟，等. 回应AIGC的信息资源管理学人思考[J]. 农业图书情报学报，2023，35（1）：4-28.

[18] 钱力，刘熠，张智雄，等. ChatGPT的技术基础分析[J]. 数据分析与知识发现，2023，7（3）：6-15.

[19] 张晓艳，王挺，陈火旺. 基于混合统计模型的汉语命名实体识别方法[J]. 计算机工程与科学，2006，28（6）：135-139.

[20] GAO W C，ZHAO S S，ZHU S Y，et al. Research on entity recognition in aerospace engine fields based on conditional random fields[J]. Journal of Physics：Conference Series，2021，1848（1）：012058.

[21] 张芳丛，秦秋莉，姜勇，等. 基于RoBERTa-WWM-BiLSTM-CRF的中文电子病历命名实体识别研究[J]. 数据分析与知识发现，2022，6（S1）：251-262.

作者简介

柳涛，男，硕士研究生，研究方向：大数据分析方法与技术。
丁陈君，女，博士，副研究员，通信作者，研究方向：生物科技战略情报研究，E-mail：dingcj@clas.ac.cn。
姜恩波，男，硕士，研究馆员，研究方向：数字图书馆平台建设。
许睿，女，硕士，助理研究员，研究方向：专利数据分析。
陈方，女，博士，研究员，研究方向：生物科技及相关领域战略情报研究。

Construction of Scholar Profile Based on Generative Pre-Trained Language Model

LIU Tao[1,2] DING ChenJun[1] JIANG EnBo[1,2] XU Rui[1] CHEN Fang[1,2]
(1. National Science Library (Chengdu), Chinese Academy of Sciences, Chengdu 610299, P. R. China; 2. Department of Information Resources Management, University of Chinese Academy of Sciences, Beijing 100190, P. R. China)

Abstract: In the era of big data, the information of scholars in the Internet that exists in a multi-source heterogeneous and unstructured form is accompanied by problems such as attribute confusion and long entities during entity extraction, which seriously affect the accuracy of the construction of scholar profiles. Meanwhile, the scholar attribute entity extraction model, as a key model in the construction of scholar profiles, still presents significant technical barriers in practical applications, which pose certain obstacles to the widespread application of scholar profiles. Therefore, based on open resources, we construct an attribute entity extraction method based on generative pre-trained language models through guided sentence modelling, autoregressive generation approach, and training corpus fine-tuning, and validate the method from four aspects: overall model effect, entity category extraction effect, instance analysis of the main influencing factors, and analysis of sample fine-tuning impact. Compared with the contrastive models, the method proposed in this paper achieves optimal performance across 12 categories of scholar attribute entities, with a comprehensive F1 score of 99.34%. It not only effectively identifies and differentiates mutually confusing attribute entities, but also enhances the extraction precision of typical long attribute entities such as “research interests” by 6.11%. This method provides more expedient and effective methodological support for the engineering application of scholar profiles.

Keywords: Generative Pre-Trained Language Model; Sample Fine-Tuning; Scholar Profile; GPT-3

（责任编辑：王玮）