· 情报方法与技术创新 ·

基于大小模型协同的情报学理论实体抽取研究

姚汝婧[1,2] 王 芳[1,2*]
(1. 南开大学商学院信息资源管理系，天津 300071；2. 南开大学网络社会治理研究中心，天津 300071)

摘 要：[目的/意义] 理论是情报学学科构建与发展中至关重要的组成部分，对理论的梳理与分析不仅有助于理解情报学学科的起源与发展脉络，也能够预测新兴技术的发展，高效和准确地识别理论实体对于促进理论研究的深化具有极为重要的作用。[方法/过程] 本文提出了一种大小模型协同的情报学理论实体抽取算法，包括词嵌入向量增强、样本难度评估和理论识别模型 3 个模块。首先利用大型语言模型对理论实体进行预识别，预识别的实体与句子中的原始词嵌入向量构成增强词嵌入向量，通过增强的词嵌入向量优化领域小模型的训练过程。此外，本文利用大模型对样本的难度进行评估，并据此调整训练策略，以提高模型性能。该算法充分结合大型语言模型强大的语义理解能力以及领域小模型的专业性。[结果/结论] 在情报学理论实体抽取数据集上展开实验，结果表明本文提出的算法有效提升了理论实体抽取的性能，在精确率、召回率、F1 指标上均实现了最优结果。

关键词：大型语言模型；情报学理论；实体识别；样本学习难度；模型协同
DOI:10.3969/j.issn.1008-0821.2025.04.001
〔中图分类号〕TP391 〔文献标识码〕A 〔文章编号〕1008-0821 (2025) 04-0003-09

Research on Information Science Theoretical Entity Extraction Based on Collaboration Between Large and Small Language Models

Yao Rujing[1,2] Wang Fang[1,2*]
(1. Department of Information Resources Management, Business School, Nankai University, Tianjin 300071, China;
2. Center for Network Society Governance, Nankai University, Tianjin 300071, China)

Abstract: [Purpose/Significance] Theory is an essential component in the construction and development of the discipline of information science. Organization and Analysis of theories not only help understand the origins and developmental trajectories of the discipline but also predict the development of emerging technologies. Efficient and accurate identification of theoretical entities plays a crucial role in deepening theoretical research. [Method/Process] This paper proposed an information science theory extraction algorithm that collaborates between large and small language models, including modules for enhanced word embedding vectors, sample difficulty assessment, and a theoretical identification model. Initially, the paper used large language models to pre-identify theoretical entities. These pre-identified entities, combined with the original word embeddings, formed the enhanced word embeddings. The training process of domain-specific small models was optimized through these enhanced word embedding vectors. Additionally, the paper used large language models to assess the difficulty of samples and adjusts training strategies accordingly to improve model performance. The proposed algorithm

收稿日期：2024-11-18
基金项目：国家社会科学基金重大项目“基于数据共享与知识复用的数字政府智能化治理研究”(项目编号：20ZDA039)。
作者简介：姚汝婧 (1995-)，女，博士研究生，研究方向：数据挖掘与知识发现。
通信作者：王芳 (1970-)，女，教授，博士生导师，研究方向：情报学理论与方法；知识发现与情感挖掘；政府信息资源管理与电子政务；政策信息分析。

fully integrated the large language models' powerful semantic understanding capabilities and the professionalism of domain-specific small models. [Result/Conclusion] Experiments conducted on a dataset for the extraction of theoretical entities in information science show that the algorithm proposed in this paper effectively improves the performance of theoretical entity extraction, achieving the best results in the metrics of precision, recall, and F1 score.

Key words: large language model; information science theory; entity recognition; sample learning difficulty; model collaboration

理论是一个学科长久发展的关键要素，在科学研究和实际应用中都有着非常重要的地位与价值[1]。情报学学科经过多年的发展，已经形成了比较完整的学科体系[2]，其中理论的构建与发展起到了核心的推动作用。对情报学理论的研究不仅可以帮助了解情报学学科的起源与发展脉络，还有助于了解未来趋势，预测和适应新兴技术的发展。当前，已经有许多学者对情报学理论进行了深入研究，涵盖了理论体系构建[3]、理论跨学科性[4-5]、理论的影响力[6]等多个方面，这些研究都凸显了理论实体本身的重要性。因此，对理论实体的识别也成为了情报学研究中的关键任务之一。

命名实体识别作为自然语言处理领域的一个重要分支，在情报学理论研究中展现出巨大的潜力，是目前情报学理论实体抽取的主要方法之一。命名实体识别技术能够自动从非结构化文本中识别和提取特定类型的实体，如人名、地名、组织机构名等，在情报学领域中，这项技术也可以被应用于识别和提取理论实体。目前，已经有大量的学者使用和设计相应的命名实体识别算法来实现理论实体的提取。例如，王昊等[1]应用 BiLSTM-CRF 深度学习模型，对 20 年来情报学领域相关文献中的情报学理论实体进行抽取。赵洪等[7]提出了一种自训练算法抽取学术文献中的理论术语。这类实体抽取模型针对特定领域进行训练和优化，具有较好的抽取性能，但是依赖于大量的标注数据，数据的质量和量级直接影响模型的表现。

近年来，大型语言模型的出现与飞速发展为情报学理论研究带来了新的机遇和挑战。以 ChatGPT 和 Llama 为代表的大语言模型展现出了卓越的自然语言理解和生成能力，为情报学理论实体抽取带来全新的范式和可能性。然而，目前对如何有效应用这些大型语言模型进行情报学理论实体抽取的探讨相对匮乏。情报学理论因其概念抽象性高、术语专业性强等特性，在抽取任务上具有很大的挑战。大型语言模型虽然在通用领域表现出色，但由于缺乏针对情报学理论的专业训练，因此在进行情报学理论实体的抽取时存在一定的局限性，比如会无法区分专业术语和通用语言之间的细微差别，理论实体抽取的准确性和可靠性有限。然而，大型语言模型强大的语义理解能力仍可以为理论实体识别任务提供有效的辅助知识。因此，针对上述问题和挑战，本文提出了一种大型语言模型与领域小模型协同的情报学理论实体抽取算法。该算法充分融合了大型语言模型强大的语言表示与理解能力，以及领域小模型的专业性，并考虑了在模型学习过程中样本的学习难度评估，从而提升了理论实体抽取任务的性能。总的来说，本文的贡献如下：

1）本文提出了一种大型语言模型与领域小模型协同的情报学理论实体抽取算法，包括 3 个模块：词嵌入向量增强、样本难度评估和理论识别模型。

2）在情报学理论实体抽取数据集上进行实验，验证了本文所提出的算法的性能。

1 相关工作

1.1 大型语言模型

随着 ChatGPT、Llama、Claude、通义千问、文心一言等大型语言模型的出现，人工智能在理解和生成自然语言方面取得了非常显著的进展。这些通用语言模型不仅推动了学术研究，也促进了各行各业应用的落地。此外，一些专门针对特定行业的大模型也相继出现，如金融领域的通义金融、ERNIE-Finance，医疗领域的通义仁心、ERNIE-Health，以及法律领域的通义法睿、ERNIE-Law 等。这些大型语言模型通过在海量文本数据上进行预训练，学习了丰富的语言特征和深层知识，在多种自然语言处理任务中展示了卓越的性能。它们的出现为研究和实际应用提供了新的视角和工具，广泛应用于自动问答、推荐系统等任务中，极大推动了行业的技术

进步和效率提升。

在自动问答任务中，程云等[8]将大模型和检索增强生成技术进行结合，构建了标准文献智能问答解决方案。Yang R 等[9]提出了一种端到端方法，使用检索增强大型语言模型进行可解释的法律问答。Li K Z 等[10]提出了一种大模型引导方法生成教育领域中可控问题。Zheng X X 等[11]提出了 KS-LLM 方法，从文档中选择有利于回答问题的知识片段，帮助大型语言模型生成更优的答案。在推荐系统中，Lubos S 等[12]介绍并评估了大型语言模型在商品推荐中提供高质量解释的能力。Wang X Y 等[13]提出了一个图注意力的大模型生成式推荐系统，将大型语言模型强大的上下文表示能力与图结构信息相结合。Kim S 等[14]提出了 A-LLMRec 系统，将大模型与协同过滤推荐系统相结合，提升推荐系统的性能。

总体而言，大型语言模型不仅在自然语言处理任务中表现出色，还通过与现有技术的结合，为解决各种复杂问题开辟了新的可能性。大语言模型凭借其强大的语义理解能力和泛化能力，能够有效处理和理解大量的非结构化数据，从而在众多领域中都有着非常广泛的应用。

1.2 学术领域的命名实体识别

命名实体识别是自然语言处理领域中的一项关键任务，其目标是从文本中识别出具有特定含义的实体，如人名、地名、机构名等。在学术领域，命名实体识别主要用于从学术文献提取关键术语实体，如从生物医学文献中提取特定的蛋白质、疾病、诊疗方法，从计算机相关文献中提取算法、数据集等。

传统的命名实体识别任务依赖于人工设计的规则或构建的词典。Zhang C 等[15]使用基于字典的方法，对自然语言处理领域的英文和中文论文中的算法提及情况进行识别和比较。Wang Y Z 等[16]手动标注了 1979—2015 年 ACL 会议论文全文中的算法实体，并对不同算法的影响力、不同时间段排名前十的算法等进行了分析。Ding R Y 等[17]利用基于字典的方法，对学术论文全文中的算法进行抽取，研究了算法引用的频率和时间的演变。Wang Y 等[18]使用字典抽取了发表在 ACL 会议上的论文全文中的算法，并对排名前十的数据挖掘算法在论文数量、提及频率和算法位置等方面进行了比较。

传统方法依赖于大量的人力物力，因此，基于机器学习和深度学习的命名实体抽取算法逐渐成为主流。Zha H W 等[19]利用深度学习算法，对表格中的算法进行提取，对算法路线图进行构建，从而描述不同算法之间的演化关系。Lei Z 等[20]将 CRF、BiLSTM+CRF 应用于学术论文模型实体的抽取中。Tuarob S 等[21]提出了一种新颖的可扩展技术对学术文档中的算法表示进行识别。Luan Y 等[22]在 LSTM-CRF 模型的基础上引入了半监督算法，提升了命名实体识别任务的性能。为了识别法律文献中的命名实体，Zhang X R 等[23]提出了一种基于 RoBERTa-GlobalPointer 的方法。章成志等[24]比较了多种机器学习算法在算法实体抽取方面的性能，并对《情报学报》期刊上 10 年的研究方法实体进行了抽取和统计分析。此外，章成志等[25]将方法细分为算法、数据集、指标和工具，利用 SciBERT-CRF 等模型对发表在 ACL 会议上的论文方法进行抽取，并分析了方法实体的演化情况。张颖怡等[26]对学术论文中问题与方法的识别和关系抽取相关研究进行了系统化的梳理。

随着 ChatGPT、Llama 等大型语言模型的兴起，利用其进行命名实体识别已成为新的研究趋势。然而，目前鲜有研究聚焦于利用大模型来抽取情报学理论实体，现有的情报学理论实体抽取的方法主要集中于基于手工标注、规则与词典、机器学习与深度学习技术。例如，刘竟等[27]人工标注了图书馆学博士学位论文中使用的理论，对图书馆学研究中的理论应用情况进行分析。王芳等[2]构建了理论识别模型对文献中的理论实体进行抽取，对情报学的理论研究形态及学术影响力等进行了分析。此外，王芳等[28]还利用词典和规则融合的方法，对情报学论文中的理论进行识别，分析了各国家(地区)情报学研究中理论应用及来源学科的特点。Zhang C 等[29]使用深度学习算法，抽取了文献中的理论实体，并对理论应用情况进行分析，揭示了情报学的跨学科性。大型语言模型具有强大的自然语言理解能力，利用大型语言模型进行情报学理论实体的抽取和分析具有广阔的应用前景和重要的研究价值。

1.3 大型语言模型在命名实体识别领域的应用

利用 ChatGPT、Llama 等大型语言模型进行命

名实体识别已成为当前研究的热点。研究者们探究了多种方法来提升大型语言模型在命名实体识别任务上的性能。Wang S 等[30]将序列标记任务转换为生成任务，并提出了一种自我验证策略，有效地提高了大型语言模型在命名实体识别任务上的性能。Xiao L 等[31]提出了一种基于大型语言模型的命名实体识别框架 LLM-DER，有效解决了特定领域中复杂结构实体识别问题。对于少样本命名实体识别，Ye J 等[32]提出了一种基于大模型的数据增强策略，提升了命名实体识别模型的性能。Santoso J 等[33]使用大模型和少量标记示例来生成命名实体识别数据，减少对大量人工注释数据的需求。Cheng Q 等[34]为利用大模型进行命名实体识别，提出了一种新的少样本提示方法。在特定领域，Peng R L 等[35]使用大模型有效地从非结构化数据中提取了农业信息。孙熠衡等[36]提出一种基于提示学习的信息抽取方法，使用 ChatGLM 大模型对标书中的领域术语进行提取。Cao M Y 等[37]分别对 ChatGLM 进行优化，使其专注于实体提取任务，在医疗领域的关键信息提取任务上实现了较好的效果。Dagdelen J 等[38]对大型语言模型进行微调，用大模型来提取材料领域的相关信息。Liu J 等[39]提出了一种基于上下文学习的抽取框架，使用大模型来对健康相关的信息进行提取。Jung S J 等[40]使用大模型来对科学文本中的蛋白质、基因和小分子进行识别。

总的来说，大型语言模型在命名实体识别领域的应用越来越广泛，并有效提升了命名实体识别任务的性能，进一步推动了命名实体识别技术的发展与应用。

2 研究方法

本研究提出了一种大小模型结合的命名实体识别算法，充分结合了大语言模型的先验知识和领域小模型的专业性，以提高理论实体识别的性能。算法整体框架如图 1 所示，主要包括 3 个模块：词嵌入向量增强、样本难度评估以及理论识别模型。具体来说，本研究利用大型语言模型对命名实体进行预识别，并对样本学习难度进行初步预评估。然后，将大模型预识别实体的词嵌入向量与领域小模型(即理论识别模型)中原始词嵌入向量结合，构成增强后的词嵌入表示，并利用大语言模型初步预评估的样本学习难度指导领域小模型初期的样本选择。

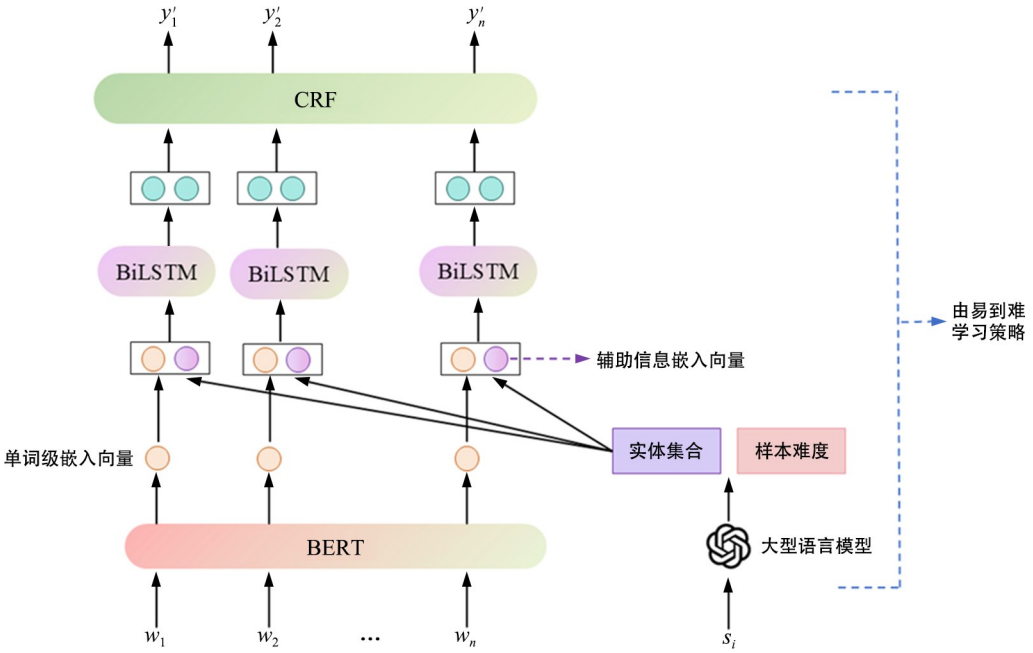

图 1 算法整体框架

Fig. 1 Framework of the Proposed Algorithm

2.1 词嵌入向量增强

本节对词嵌入向量增强模块进行详细介绍。首先利用大语言模型来初步识别理论实体，然后将预识别的理论实体作为辅助信息，与原始词嵌入向量结合，构成增强后的词嵌入表示。

具体来说，对于每一个输入句子 $s_i = \{w_{i1}, w_{i2},$

…，$w_{in}\}$，s_i 表示第 i 个句子，w_{ij}为第 i 个句子的第 j 个词。首先利用上下文学习(In-Context Learning)技术，指导大模型识别句子 s_i 的理论实体，如式（1）所示：

$$E_i = LLM(s_i) \tag{1}$$

其中，E_i 表示大语言模型对句子 s_i 识别出的理论实体集合，LLM 表示大型语言模型。

对于句子 s_i 中的一个词 w_{ij}，如果其在大模型识别出的实体集 E_i 中，则为其分配一个可学习的嵌入向量 v_{ij}，$v_{ij} \in R^d$。如果其不在 E_i 中，则为其分配一个维 d 的 0 向量。为了充分结合大模型和小模型各自的能力，将上述嵌入向量与小模型对该词的嵌入向量 r_{ij}进行拼接，得到增强后的嵌入向量 u_{ij}，如式（2）所示：

$$\begin{aligned} &u_{ij} = r_{ij} \oplus v_{ij}，\text{如果 } w_{ij} \in E_i \\ &u_{ij} = r_{ij} \oplus 0，\text{如果 } w_{ij} \notin E_i \end{aligned} \tag{2}$$

2.2 样本难度评估

在深度学习模型的训练过程中，理解和适应不同样本的学习难度对模型性能的提升至关重要。模仿人类学习由浅入深的特点，Kumar M P 等[41]提出了自步学习。该策略按照从简单到复杂的顺序逐步训练模型，从而提高模型的学习效率和性能。自步学习的优化方式如式（3）所示：

$$\min \sum_i v_i l_i \tag{3}$$

其中，l_i 表示句子 s_i 的损失，v_i 表示句子 s_i 的权重，v_i 的取值如式（4）所示：

$$v_i = \begin{cases} 1, \text{ if } \quad l_i < \lambda \\ 0, \text{ if } \quad l_i \geqslant \lambda \end{cases} \tag{4}$$

其中，λ 为超参数，随着训练轮次的增加，λ 值逐渐增大，越来越多的样本被纳入训练过程中。

自步学习已被证实在多种任务上都可以有效提升模型的性能[42-43]，然而，在模型训练初期，模型可能倾向于选择它认为学习难度低的样本。鉴于早期模型的性能有限，其对样本学习难度的评估可能存在较大偏差，为了克服这一问题，引入了大模型来进行样本学习难度的初步预评估。利用大模型强大的泛化能力和特征提取能力，更准确判断样本难度，从而指导初期的样本选择。随着模型训练的进行，特别是进入中后期阶段，本文开始根据损失函数的反馈调整对样本学习难度的评估。这种动态调整策略能够根据模型的实时学习状态优化训练过程，确保模型能够有效学习到更复杂的样本，全面提升模型的性能。

具体来说，对于一个问题 s_i，首先利用大模型对 s_i 中的理论实体重复进行 k 次识别，每次实体识别的结果记为 E_i^k，E_i^k 表示第 i 个句子在第 k 次实验中识别出的理论实体集合。对于每次实验，其相应的预测标签序列为 y_i^k，y_i^k 表示第 i 个句子的第 k 次标签预测结果。$y_i^k = \{y_{i1}^k, \cdots, y_{ij}^k, \cdots, y_{in}^k\}$，$y_{ij}^k$表示在第 k 次实验中，第 i 个句子的第 j 个词的标签预测结果。

熵是信息论中用于量化不确定性的重要方式，在本文中通过计算每个句子预测标签的熵来实现样本难度的评估，熵越大表明模型对该样本的预测难度越高，学习难度越大。具体来说，对于句子 s_i 中的每个词 w_{ij}，首先计算每个可能的标签在 k 次预测中出现的频率，如式（5）所示：

$$p_{ij}^l = \frac{T_{ij}^l}{k} \tag{5}$$

其中，T_{ij}^l为标签 l 在 k 次预测中出现的次数。

然后，使用式（6）计算每个词的熵：

$$H_{ij} = -\sum_l p_{ij}^l \log(p_{ij}^l) \tag{6}$$

为了量化整个句子的难度，将句子中所有词的熵的平均值作为句子整体的熵，这个平均熵值用来衡量整个句子的难度，如式（7）所示：

$$H_i = \frac{1}{n} \sum_{j=1}^{n} H_{ij} \tag{7}$$

在模型训练初期，即当前 Epoch 小于阈值 γ 时，采用式（8）来计算权重 v_i：

$$v_i = \begin{cases} 1, \text{ if } \quad H_i < \tau \\ 0, \text{ if } \quad H_i \geqslant \tau \end{cases} \tag{8}$$

随着模型训练的进行，采用式（9）来计算权重：

$$v_i = \begin{cases} 1, \text{ if } \quad l_i < \lambda \\ 0, \text{ if } \quad l_i \geqslant \lambda \end{cases} \tag{9}$$

2.3 理论识别模型

本文采用经典的命名实体识别模型来作为小模型基础网络架构(即理论识别模型)，以 BERT-BiLSTM-CRF 为例，对于一个句子 $s_i = \{w_{i1}, w_{i2}, \cdots, w_{in}\}$，$s_i$ 表示第 i 个句子，w_{ij}为第 i 个句子的第 j 个词。首先利用 *BERT* 来得到每个词的嵌入表示，如

式（10）所示：

$$r_{ij}=BERT(w_{ij}) \tag{10}$$

然后利用 2.1 节中的方法，得到增强后的嵌入向量 u_{ij}。为了得到整个句子的增强嵌入向量，将句子中所有词的嵌入取平均值来作为句子的嵌入，如式（11）所示：

$$U_i=\frac{1}{n}\sum_{j=1}^{n}u_{ij} \tag{11}$$

BiLSTM 能够有效捕获到正向和反向的上下文信息，将增强后的嵌入向量输入到 BiLSTM 中，如式（12）~（14）所示：

$$\vec{h}_i=LSTM(\vec{h}_{i-1},U_i) \tag{12}$$

$$\overleftarrow{h}_i=LSTM(\overleftarrow{h}_{i-1},U_i) \tag{13}$$

$$h_i=\vec{h}_i\oplus\overleftarrow{h}_i \tag{14}$$

CRF 模块能够考虑相邻标签的依赖性，非常好地适用于命名实体识别任务，将得到的隐藏层向量 h_i 输入到 CRF 模块中来得到最后的预测标签。在整个模型的训练过程中，使用 v_i 来控制参与训练的样本，损失函数的计算如式（15）所示：

$$L=-\sum_{i=1}^{N}v_i\cdot l_i \tag{15}$$

其中，l_i 表示样本 s_i 的损失，v_i 表示样本 s_i 的权重。

3 实 验

3.1 数据集

Zhang C 等[29]的研究以 2001—2021 年发表在《Journal of the Association Society for Information Science and Technology》《Information Processing and Management》《Journal of Documentation》《Journal of Informetrics》《Journal of Information Science》《Information and Management》《Journal of Librarianship and Information Science》《Library and Information Science Research》8 本国际情报学期刊的论文为数据来源，对其中的6 000条句子进行了标注。在此基础上，为进一步提高训练集标注的准确性，本研究重新招募了 10 名情报学专业的硕士生和博士生，复核上述标注的句子中所包含的理论实体并补充标注，最终共完成3 367条数据的标注。本文按照8∶1∶1的比例将数据集随机划分为训练集、验证集和测试集，数据集详细情况如表 1 所示。

表 1 数据集信息

Tab. 1 Dataset Information

数据集划分	数量
训练集	2 693
验证集	337
测试集	337

3.2 对比算法和参数设置

为了全面评估本研究所提出算法的性能，且保证实验的可重复性，本文采用开源的 Llama-2-13B 和 Llama-3-8B 两种大模型进行实验。这两个模型均由 Meta 发布，Llama-2-13B 的参数量为 130 亿，Llama-3-8B 的参数量为 80 亿，二者在多种自然语言处理任务上展现出优越的性能。对于小模型的性能对比，本文采用以下命名实体识别算法作为基准：BiLSTM－CRF[44]、BERT－CRF[45]、SciBERT－CRF[46]、BERT－BiLSTM－CRF[47]和 SciBERT－BiLSTM－CRF[48]。

对于 Llama-2-13B 和 Llama-3-8B，本文的温度值设为 0.8，Top-p 值为 0.9，最大 Token 限制为4 096。对于 BiLSTM-CRF 算法，本文采用 300 维的 Glove 词向量，隐藏层维度设为 256，Epoch 设为 50。对于 BERT-CRF、SciBERT-CRF、BERT-BiLSTM-CRF、SciBERT－BiLSTM－CRF，Epoch 设为 10，学习率设为 1e-5，BiLSTM 隐藏层维度设为 256。本文提出的算法中，大模型对理论实体的识别次数 k 设为 3，辅助嵌入向量维度设为 300，阈值 γ 设为 Epoch/3，τ 设为所有样本熵值的平均值。

3.3 评价指标

本文采用命名实体识别任务中常用的指标来衡量模型的性能，包括精确率、召回率和 F1。精确率指的是实际为实体且模型预测也为实体的标签数量占所有被预测为实体的标签数量的比例，计算如式（16）所示：

$$Precision=\frac{TP}{TP+FP} \tag{16}$$

其中，TP 为真正例(实际为实体且模型预测也为实体)，FP 为假正例(实际不是实体但模型预测为实体)。

召回率指的是实际为实体且模型预测也为实体的标签数量占所有实际为实体的标签数量的比例，

计算如式（17）所示：

$$Recall=\frac{TP}{TP+FN} \tag{17}$$

其中，FN 为假负例（实际为实体但模型预测不是实体）。

$F1$ 指的是精确率与召回率的调和平均数，计算如式（18）所示：

$$F1=2\cdot\frac{Precision\cdot Recall}{Precision+Recall} \tag{18}$$

3.4 实验结果

表 2 为使用 Llama-2-13B 大模型时各个算法在理论实体识别任务上的实验结果，表 3 为使用 Llama-3-8B 大模型时，各个算法在理论实体识别任务上的实验结果。

表 2 使用 Llama-2-13B 大模型时算法结果

Tab. 2 Algorithm Results When Using the Llama-2-13B

模 型	精确率	召回率	F1
BiLSTM-CRF	0.7067	0.6607	0.6829
BERT-CRF	0.6947	0.7416	0.7174
SciBERT-CRF	0.7412	0.7528	0.7469
BERT-BiLSTM-CRF	0.7171	0.7348	0.7259
SCiBERT-BiLSTM-CRF	0.7294	0.7753	0.7516
Our Model+BiLSTM-CRF	0.7115	0.6726	0.6915
Our Model+BERT-CRF	0.7254	0.7355	0.7304
Our Model+SciBERT-CRF	0.7525	0.7539	0.7532
Our Model+BERT-BiLSTM-CRF	0.7226	0.7489	0.7355
Our Model+SciBERT-BiLSTM-CRF	0.7370	0.7764	0.7562

表 3 使用 Llama-3-8B 大模型时算法结果

Tab. 3 Algorithm Results When Using the Llama-3-8B

模 型	精确率	召回率	F1
BiLSTM-CRF	0.7067	0.6607	0.6829
BERT-CRF	0.6947	0.7416	0.7174
SciBERT-CRF	0.7412	0.7528	0.7469
BERT-BiLSTM-CRF	0.7171	0.7348	0.7259
SCiBERT-BiLSTM-CRF	0.7294	0.7753	0.7516
Our Model+BiLSTM-CRF	0.7121	0.6755	0.6933
Our Model+BERT-CRF	0.7263	0.7393	0.7327
Our Model+SciBERT-CRF	0.7538	0.7552	0.7545
Our Model+BERT-BiLSTM-CRF	0.7239	0.7487	0.7361
Our Model+SciBERT-BiLSTM-CRF	0.7381	0.7763	0.7567

由表 2 可以看出，在使用 Llama-2-13B 大模型时，本文提出的模型在多组对比实验中均表现出优异的性能，在精确率、召回率、F1 这 3 个指标上均实现了最佳结果。当将本文的模型分别与 BiLSTM-CRF、BERT-CRF、SciBERT-CRF、BERT-BiLSTM-CRF、SCiBERT-BiLSTM-CRF 相结合后，相较于单独使用这些模型，F1 值均有不同程度的提升，分别提升了 0.86%、1.30%、0.63%、0.96%、0.46%。当本文的模型与 SciBERT-CRF 结合后，其 F1 值达到了 75.32%，在与 SciBERT-BiLSTM-CRF 结合后，F1 值进一步提高至 75.62%，实现了最佳性能。

由表 3 可以看出，在使用 Llama-3-8B 大模型时，本文提出的模型在多组对比实验中也都表现出了优异的性能。当本文提出的框架与基线模型结合时，无论是 BiLSTM-CRF、BERT-CRF、SciBERT-CRF，还是 BERT-CRF、SciBERT-CRF 与 BiLSTM 的组合，在性能方面均有进一步提升。具体来说，与只使用 BiLSTM-CRF、BERT-CRF、SciBERT-CRF、BERT-LSTM-CRF、SCiBERT-LSTM-CRF 相比，使用本文的模型以后，F1 值分别提升了 1.04%、1.53%、0.76%、1.02%、0.51%。通过上述实验结果可以看出，本文提出的算法有效提升了情报学理论实体的抽取性能。

此外，通过实验结果可以看出，无论是结合传统的 BiLSTM-CRF 结构，还是融合 BERT-CRF、SciBERT-CRF、BERT-BiLSTM-CRF、SciBERT-BiLSTM-CRF 的多层结构，本文提出的模型在各种组合下均能带来显著的性能提升，这表明本文的模型具有较好的适应性与通用性，能够在不同小模型和大模型组合下稳定发挥作用。

为了验证本文算法中各个模块的性能，在使用 Llama-3-8B 作为大模型、BERT-LSTM-CRF 作为小模型时进行了消融实验，结果如表 4 所示。相比于本文的整体模型，不融入词嵌入向量增强模块的模型在精确率、召回率、F1 值 3 个指标上分别下降了 0.42%、0.79%、0.60%。相比于本文的整体模型，不融入样本难度模块的模型在精确率、召回率、F1 值 3 个指标上分别下降了 0.53%、0.59%、0.56%。由消融实验可以看出，词嵌入向量增强模块和样本

难度模块均起到了积极作用，在词嵌入向量增强模块和样本难度模块同时使用时，整体模型实现了最佳性能。词嵌入向量增强模块能够加强词汇层面的特征表示，使模型能更精确地捕捉特定领域词汇的特征，样本难度模块从训练样本的难易程度出发，动态调整训练模型的样本。二者结合后，模型既拥有较强的词汇表征能力，又具备调整训练策略的机制，从而实现了性能的全面提升。

表 4 消融实验

Tab. 4 Ablation Study

模 型	精确率	召回率	F1
Our Model+BERT-BiLSTM-CRF	0.7239	0.7487	0.7361
Our Model w/o 词嵌入向量增强	0.7197	0.7408	0.7301
Our Model w/o 样本难度	0.7186	0.7428	0.7305
BERT-BiLSTM-CRF	0.7171	0.7348	0.7259

此外，本研究对只使用大模型进行初步理论实体识别的识别质量进行了分析，结果如表 5 所示。采用 Llama-2-13B 作为大模型时，在测试集上，精确率、召回率、F1 分别为 0.4454、0.5629、0.4973。采用 Llama-3-8B 作为大模型时，在测试集上，精确率、召回率、F1 分别为 0.4613、0.5783、0.5132。由此可见，只利用大型语言模型进行情报学理论实体识别的性能欠佳。这可能是由于大型语言模型虽然具有强大的语言理解能力和广泛的知识，但是其在特定领域(如情报学领域)的专业知识不够充分，导致其在该领域理论实体抽取性能相对不足。

表 5 大型语言模型理论实体识别性能

Tab. 5 Performance of LLMs in Theory Entity Recognition

模 型	精确率	召回率	F1
Llama-2-13B	0.4454	0.5629	0.4973
Llama-3-8B	0.4613	0.5783	0.5132

4 总结与展望

本文提出了一种结合大型语言模型与领域小模型的算法来对情报学理论实体进行抽取。该算法包括词嵌入向量增强、样本难度评估和理论识别模型 3 个模块。首先，使用大型语言模型对理论实体进行预识别，将预识别的结果作为辅助词嵌入向量，与原始词嵌入向量进行结合，增强词嵌入表示。此外，考虑到在训练初期，模型对样本难度的评估存在较大误差，本文利用大型语言模型的预测结果计算熵值从而评估样本的难度，据此调整训练策略，以提高模型性能。本文将词嵌入向量增强模块和样本难度评估模块与经典的命名实体识别算法进行结合，充分利用了大型语言模型强大的语言表示与理解能力，以及领域小模型的专业性。在情报学理论实体数据集上进行实验，结果表明本文提出的算法有效提升了理论实体抽取的性能，在精确率、召回率、F1 指标上均实现了最优结果。

在未来的研究中，将考虑如何更好地利用大型语言模型的能力来帮助优化领域小模型，并将探索结合文本以外的其他数据类型来进行理论实体抽取，如图像、声音和视频等。此外，还考虑将本文提出的算法应用于其他领域，如医学、法律和金融领域等。

参 考 文 献

[1] 王昊，邓三鸿，苏新宁，等. 基于深度学习的情报学理论及方法术语识别研究 [J]. 情报学报，2020，39 (8)：817-828.

[2] 王芳，赵洪，张维冲. 我国情报学科理论研究形态及学术影响力的全数据分析 [J]. 图书情报知识，2018，35 (6)：15-28.

[3] 马海群，张斌. 基于文献计量的中国特色情报学理论体系构建研究 [J]. 数字图书馆论坛，2022，(10)：49-55.

[4] 张海，石晶，孔晔晗. 图书情报学理论方法跨学科影响力成因要素研究 [J]. 情报理论与实践，2024，47 (4)：67-74.

[5] 杨志刚，郑禧溁，刘竟，等. 图书情报学研究方法和理论的跨学科使用意愿联动效应研究 [J]. 图书情报工作，2023，67 (15)：94-104.

[6] Wang F，Wang X Y. Tracing Theory Diffusion：A Text Mining and Citation-Based Analysis of TAM [J]. Journal of Documentation，2020，76 (6)：1109-1134.

[7] 赵洪，王芳. 理论术语抽取的深度学习模型及自训练算法研究 [J]. 情报学报，2018，37 (9)：923-938.

[8] 程云，吕爽，陈国祥. 基于大模型的标准文献智能问答技术研究 [J]. 信息技术与标准化，2024，(8)：38-43.

[9] Yang R，Yang B M，Ouyang S X，et al. Leveraging Large Language Models for Concept Graph Recovery and Question Answering in NLP Education [EB/OL]. https://arxiv.org/pdf/2402.14293，2024-02-22.

[10] Li K Z，Zhang Y. Planning First，Question Second：An LLM-Guided Method for Controllable Question Generation [C] //Findings of the Association for Computational Linguistics ACL 2024，2024：4715-4729.

[11] Zheng X X, Che F H, Wu J Y, et al. KS-LLM: Knowledge Selection of Large Language Models with Evidence Document for Question Answering [EB/OL]. https://arxiv.org/abs/2404.15660, 2024-04-24.

[12] Lubos S, Tran T N T, Felfernig A, et al. LLM-Generated Explanations for Recommender Systems [C] //Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, 2024: 276-285.

[13] Wang X Y, Wu L, Hong L J, et al. LLM-Enhanced User-Item Interactions: Leveraging Edge Information for Optimized Recommendations [EB/OL]. https://arxiv.org/abs/2402.09617, 2024-02-14.

[14] Kim S, Kang H, Choi S, et al. Large Language Models Meet Collaborative Filtering: An Efficient All-Round LLM-Based Recommender System [C] //Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024: 1395-1406.

[15] Zhang C, Ding R, Wang Y. Algorithms Mention in Full-Text Content of Article from NLP Domain: Comparative Analysis between English and Chinese [J]. Data Science and Informetrics, 2021, 1 (2): 19-33.

[16] Wang Y Z, Zhang C Z. Using the Full-Text Content of Academic Articles to Identify and Evaluate Algorithm Entities in the Domain of Natural Language Processing [J]. Journal of Informetrics, 2020, 14 (4): 101091.

[17] Ding R Y, Wang Y Z, Zhang C Z. Investigating Citation of Algorithm in Full-Text of Academic Articles in NLP Domain: A Preliminary Study [C] //Proceedings of the 17th International Conference on Scientometrics and Informetrics(ISSI 2019), Rome, Italy, 2019: 2726-2728.

[18] Wang Y, Zhang C. Using Full-Text of Research Articles to Analyze Academic Impact of Algorithms [C] //International Conference on Information. Springer, Cham, 2018: 395-401.

[19] Zha H W, Chen W H, Li K Q, et al. Mining Algorithm Roadmap in Scientific Publications [C] //Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019: 1083-1092.

[20] Lei Z, Wang D. Model Entity Extraction in Academic Full Text Based on Deep Learning [C] //Proceedings of the 17th International Conference on Scientometrics and Informetrics(ISSI), Vol Ⅱ, pp. 2732-2733.

[21] Tuarob S, Bhatia S, Mitra P, et al. AlgorithmSeer: A System for Extracting and Searching for Algorithms in Scholarly Big Data [J]. IEEE Transactions on Big Data, 2016, 2 (1): 3-17.

[22] Luan Y, Ostendorf M, Hajishirzi H. Scientific Information Extraction with Semi-SSupervised Neural Tagging [C] //Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 2641-2651.

[23] Zhang X R, Luo X D, Wu J Y. A RoBERTa-GlobalPointer-Based Method for Named Entity Recognition of Legal Documents [C] //2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia. IEEE, 2023: 1-8.

[24] 章成志, 张颖怡. 基于学术论文全文的研究方法实体自动识别研究 [J]. 情报学报, 2020, 39 (6): 589-600.

[25] 章成志, 谢雨欣, 张恒. 学术文献全文内容中的方法实体细粒度抽取及演化分析研究 [J]. 情报学报, 2023, 42 (8): 952-966.

[26] 张颖怡, 章成志, He D Q. 学术论文中问题与方法识别及其关系抽取研究综述 [J]. 图书情报工作, 2022, 66 (12): 125-138.

[27] 刘竟, 王亚楠, 杨志刚. 我国图书馆学博士学位论文的理论应用分析 [J]. 图书情报工作, 2021, 65 (3): 34-42.

[28] 王芳, 杨京, 陈锋. 情报学研究中理论应用的国际比较 [J]. 情报学报, 2018, 37 (12): 1262-1274.

[29] Zhang C, Wang F, Huang Y, et al. Interdisciplinarity of Information Science: An Evolutionary Perspective of Theory Application [J]. Journal of Documentation, 2024, 80 (2): 392-426.

[30] Wang S, Sun X, Li X, et al. Gpt-Ner: Named Entity Recognition via Large Language Models [EB/OL]. https://arxiv.org/abs/2304.10428, 2023-10-07.

[31] Xiao L, Xu Y, Zhao J. LLM-DER: A Named Entity Recognition Method Based on Large Language Models for Chinese Coal Chemical Domain [EB/OL]. https://arxiv.org/abs/2409.10077, 2024-09-16.

[32] Ye J, Xu N, Wang Y, et al. Llm-da: Data Augmentation via Large Language Models for Few-Shot Named Entity Recognition [EB/OL]. https://arxiv.org/abs/2402.14568, 2024-02-22.

[33] Santoso J, Sutanto P, Cahyadi B, et al. Pushing the Limits of Low-Resource NER Using LLM Artificial Data Generation [C] // Findings of the Association for Computational Linguistics ACL 2024, 2024: 9652-9667.

[34] Cheng Q, Chen L Q, Hu Z X, et al. A Novel Prompting Method for Few-Shot NER via LLMS [J]. Natural Language Processing Journal, 2024, 8: 100099.

[35] Peng R L, Liu K, Yang P, et al. Embedding-Based Retrieval with LLM for Effective Agriculture Information Extracting from Unstructured Data [EB/OL]. https://arxiv.org/abs/2308.03107, 2023-08-06.

[36] 孙熠衡, 刘茂福. 基于知识提示微调的标书信息抽取方法 [J/OL]. 计算机应用: 1-10 [2025-01-16]. http://kns.cnki.net/kcms/detail/51.1307.TP.20240828.1008.002.html.

[37] Cao M Y, Wang H, Liu X M, et al. LLM Collaboration PLM Improves Critical Information Extraction Tasks in Medical Articles [C] //China Health Information Processing Conference. Singapore: Springer Nature Singapore, 2023: 178-185.

(下转第 73 页)

Based on the Integrated Fear Acquisition Theory [J]. Technology in Society, 2021, 63 (4): 101410.

[64] Dawid B, Brenda C S. Barriers to Adopting Automated Organisational Decision-Making Through the Use of Artificial Intelligence [J]. Management Research Review, 2024, 47 (1): 64-85.

[65] Schweitzer F, Belk R, Jordan W, et al. Servant, Friend or Master? The Relationships Users Build with Voice-Controlled Smart Devices [J]. Journal of Marketing Management, 2019, 35 (7-8): 693-715.

[66] Park C S, Kaye B K. Smartphone and Self-Extension: Functionally, Anthropomorphically, and Ontologically Extending Self via the Smartphone [J]. Mobile Media & Communication, 2019, 7 (2): 215-231.

[67] Petriglieri L J. Under Threat: Responses to and the Consequences of Threats to Individuals Identities [J]. Academy of Management Review, 2011, 36 (4): 641-662.

[68] Nach H, Lejeune A. Coping with Information Technology Challenges to Identity: A Theoretical Framework [J]. Computers in Human Behavior, 2010, 26 (4): 618-629.

[69] Mirbabaie M, Stieglitz S, Brünker F, et al. Understanding Collaboration with Virtual Assistants: The Role of Social Identity and the Extended Self [J]. Business & Information Systems Engineering, 2020, 63 (1): 21-37.

[70] Precedence Research. Artificial Intelligence (AI) Market [EB/OL]. https://www.precedence-research.com/artificial-intelligence-market/, 2024-01-06.

[71] Questmobile. 2024 生成式 AI 及 AIGC 应用洞察报告: 头部 App 应用去重月活用户突破 5000 万, C 端、B 端机会蜂拥而至 [EB/OL]. https://www.questmobile.com.cn/research/report/1767395734913650690/, 2024-03-12.

[72] Kautish P, Khare A. Investigating the Moderating Role of AI-Enabled Services on Flow and Awe Experience [J]. International Journal of Information Management, 2022, 66 (5): 102519.

[73] Johnson D G, Verdicchio M. AI Anxiety [J]. Journal of the Association for Information Science and Technology, 2017, 68 (9): 2267-2270.

[74] Dai B, Ali A, Wang H. Exploring Information Avoidance Intention of Social Media Users: A Cognition-Affect-Conation Perspective [J]. Internet Research, 2020, 30 (5): 1455-1478.

[75] Cezar B G D S, Maçada A C G. Cognitive Overload, Anxiety, Cognitive Fatigue, Avoidance Behavior and Data Literacy in Big Data Environments [J]. Information Processing & Management, 2023, 60 (6): 103482.

(责任编辑: 郭沫含)

(上接第 11 页)

[38] Dagdelen J, Dunn A, Lee S, et al. Structured Information Extraction from Scientific Text with Large Language Models [J]. Nature Communications, 2024, 15 (1): 1418.

[39] Liu J, Wang J, Huang H, et al. Improving LLM-Based Health Information Extraction with In-Context Learning [C] //China Health Information Processing Conference. Singapore: Springer Nature Singapore, 2023: 49-59.

[40] Jung S J, Kim H, Jang K S. LLM Based Biological Named Entity Recognition from Scientific Literature [C] //2024 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE, 2024: 433-435.

[41] Kumar M P, Packer B, Koller D, et al. Self-Paced Learning for Latent Variable Models [C] //Proceedings of the 24th International Conference on Neural Information Processing Systems, 2010: 1189-1197.

[42] 张宇, 刘波. 基于自步学习策略的归纳式迁移学习模型研究 [J]. 广东工业大学学报, 2023, 40 (4): 31-36.

[43] 江雨燕, 陶承凤, 李平. 数据增强和自适应自步学习的深度子空间聚类算法 [J]. 计算机工程, 2023, 49 (8): 96-103, 110.

[44] Huang Z H, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging [EB/OL]. https://arxiv.org/abs/1508.01991, 2015-08-09.

[45] Souza F, Nogueira R, Lotufo R. Portuguese Named Entity Recognition using BERT-CRF [EB/OL]. https://arxiv.org/abs/1909.10649, 2020-02-27.

[46] Lopez P, Du C F, Cohoon J, et al. Mining Software Entities in Scientific Literature: Document-level NEW for an Extremely Imbalance and Large-scale Task [C] //Proceedings of the 30th ACM International Conference on Information & Knowledge Management, 2021: 3986-3995.

[47] Dai Z J, Wang X T, Ni P, et al. Named Entity Recognition Using BERT BiLSTM CRF for Chinese Electronic Health Records [C] //2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). IEEE, 2019: 1-5.

[48] Zhang H, Zhang C Z, Wang Y Z. Revealing the Technology Development of Natural Language Processing: A Scientific Entity-Centric Perspective [J]. Information Processing & Management, 2024, 61 (1): 103574.

(责任编辑: 郭沫含)