doi:10.3772/j.issn.2095-915x.2025.03.007

面向开源情报的实体链接技术研究综述

韩雪[1] 陈秋琼[1] 戴耀[1] 胡佳荣[2]

1. 海军大连舰艇学院 大连 116000；
2. 大连海事大学信息科学技术学院 大连 116000

开放科学（资源服务）标识码（OSID）

摘要：[目的 / 意义] 开源情报已成为核心研究论题，其战略价值在于为决策者在复杂环境下提供关键信息支持，以辅助制定最优决策。[方法 / 过程] 详细阐述开源情报实体链接各模块运行流程，整理实体链接技术发展脉络，并结合最新科技产物 DeepSeek 探讨其对开源情报实体链接领域的影响。[结果 / 结论] 通过梳理表明现有技术理论上能有效解决开源情报实体链接问题，但随着时代发展，多模态、小样本等问题接踵而至，因此仍需全新实体链接相关技术破旧立新。

关键词：开源情报；实体链接；实体消歧；关系抽取；DeepSeek

中图分类号：G35

A Comprehensive Survey on Entity Linking Techniques Tailored for Open Source Intelligence

HAN Xue[1] CHEN Qiuqiong[1] DAI Yao[1] HU Jiarong[2]

1. Dalian Naval Academy, Dalian 116000, China;
2. Information Science and Technology College, Dalian Maritime University, Dalian 116000, China

Abstract: [Objective/Significance] Open source intelligence has become a pivotal research focus, with its strategic significance rooted in furnishing decision-makers with essential information support within intricate environments to aid in formulating optimal decisions. [Methods/Processes] This paper elaborates on the operational workflows of various modules within OSINT entity linking, organizes the developmental trajectory of entity linking technologies, and examines the impact of the latest technological innovation, DeepSeek, on the field of OSINT entity linking. [Results/Conclusions] Through a comprehensive review, it is evident that existing technologies can theoretically effectively address the issues in OSINT entity linking. However, with the advancement of time, new challenges such as multimodality and small sample sizes have emerged one after another. Therefore, there is still a need for innovative entity linking technologies to break with tradition and establish new paradigms.

Keywords: Open Source Intelligence; Entity Linking; Entity Disambiguation; Relation Extraction; DeepSeek

作者简介 韩雪（1997-），硕士，助教，主要研究方向为人工智能、开源情报等；陈秋琼（1987-），硕士，副教授，主要研究方向为网络对抗、开源情报等，E-mail：cqq0556246314@126.com；戴耀（1978-），博士，副教授，主要研究方向为人工智能、无人作战等；胡佳荣（2002-），硕士研究生，主要研究方向为人工智能、自然语言、计算机图形学等。

引用格式 韩雪，陈秋琼，戴耀，等 . 面向开源情报的实体链接技术研究综述 [J]. 情报工程，2025, 11(3): 78-91.

引言

开源情报[1]（Open Source Intelligence，OSINT）作为信息社会的重要资源，凭借低风险、低成本、高收益的优势，通过信息抽取模型反复进行提炼，被广泛应用于军事安全、政府决策、企业竞争、公共管理等方面。然而，由于开源情报来源具有广泛性、不确定性，格式具有异构性、动态性，实体具有多样性和歧义性等缺陷，如何提取和有效利用实体信息成为情报分析领域所面临的重要挑战。

现有开源情报处理方向综述大多聚焦于网络威胁和知识提取，反映出在处理开源情报时更侧重于解决动态性、误导性、异构性问题，而忽略广泛性、多样性、歧义性问题。实体链接[2]（Entity Linking，EL）作为自然语言处理领域的一项重要技术，聚焦开源情报实体本身存在的问题，结合知识库进行准确匹配，在实现信息结构化表示和关联分析的基础之上，为后续开源情报方面的信息抽取、语义理解、知识库扩容、问答系统生成和知识图谱构建等领域工作提供决定性帮助。从应用层面分析，实体链接技术无疑是解决上述问题的最优方案，但实体链接领域缺乏对开源情报方面的文献汇总并形成完整体系框架。因此本文将对该技术进行研究，按照实体链接结构梳理各模块的发展脉络，为开源情报领域提供一定的参考。

1 模型建构

实体链接是指将信息中所提及的命名实体映射到知识库（KB）中对应实体的过程，其中信息中所提及的命名实体被称为实体提及[3]（mentions），知识库中所对应的实体称为实体概念（concepts）。值得注意的是，实体提及真实存在于现实世界，而实体概念只是已经规范好的概念性数据集，是为实体提及提供辅助支撑的外部资源。

开源情报实体链接技术旨在在实体链接的基础之上，将信息范围锁定在公开状态，且任何人都能通过分析获得敌方政治、军事、经济、科技等方面内容，利用预先标识映射到情报知识库，抽取并融合得到相关实体和关系。换言之，开源情报实体链接技术本质上是具有特定指向信息的实体链接，因此要结合开源情报自身特点，有针对性地选择合适的实体链接技术。

要选择合适的开源情报实体链接技术，首先要知晓该技术为综合性自然语言处理模型，这一点可从模型架构中窥探出来。Shen 等[4]曾指出实体链接基础过程要经过命名实体识别[5]（Named Entity Recognition，NER）、候选实体与候选排序，如图 1 所示。

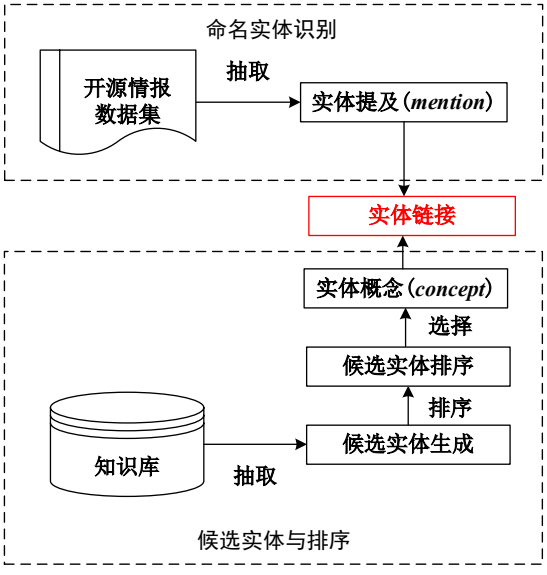

图 1　基础开源情报实体链接流程图

doi:10.3772/j.issn.2095-915x.2025.03.007

若对自然语言处理有些许了解，可发现实体抽取（实体识别的一部分）、实体排序（实体处理的一部分）等为独立自然语言处理任务，即可成为单独一门学科进行深入研究，而实体链接依托其他部分自然语言处理任务才能工作，可以说该技术起到“中转站”的作用，因此在系统研究开源情报实体链接技术时，一是不能脱离开源情报特性，二是要结合其他自然语言处理任务技术进行整合。

1.1 命名实体识别

在实体链接之前，第一步就需抽取实体提及，这一步是命名实体识别技术的核心，可以说选择适合开源情报的命名实体识别技术等同于提高实体链接的准确率。

1.1.1 基于专家领域的规则方法

早期命名实体识别采用基于专家领域的规则方法，该方法依赖于专家在某一领域所预定义的规则来进行模式匹配和识别文本中命名实体，例如在某军事领域开源情报中存在大量可疑电话号码，可对区号、长度等进行限定，生成仅识别可疑电话号码的正则表达式。这样做的好处是不需要大量被标记好的训练样本也能达到高准确率与高识别速度，但该方法极其依赖于专家所制定的规则质量，如果专家对可疑电话号码判断存有误差，那所获得的实体质量也会受影响。除此之外，若换为经济领域开源情报，仍使用该正则表达式可能会无法识别出实体，换言之在跨领域上存在局限性，因此该方法需要高昂的人工成本来定制各领域高质量规则。由于专家领域规则方法的特性，目前只有领域文本处理、实时应用等特定场景下坚持使用基于规则的命名实体识别方法，比如GATE 开源系统 [6] 的 ANNIE 插件就是通过定义一系列规则识别特定领域文本中命名实体；FACILE 则是另一个基于规则的命名实体识别系统，它参加了 MUC（Message Understanding Conference）评测，并取得了较好的成绩。

早期关于开源情报的命名实体识别虽常用于军事安全领域，但网络的发展使来自各组织、各地域、各机构的成员在科技情报、商业竞争、公共管理等领域都留下海量信息痕迹，导致各领域的开源情报盘根错节，仅依靠单领域专家不能完全处理海量开源情报。因此在 1990 年，也就是机器学习技术兴起之时，基于大规模语料库的统计方法已逐渐替代专家领域规则，成为最主流的命名实体识别技术。

1.1.2 基于大规模语料库的统计方法

基于大规模语料库的统计方法主要专注于解决上述所存在的缺陷，通过机器学习技术并结合数据自身统计规律进行建模，这意味着该方法减轻人工标注依赖，增强跨领域迁移性。该命名实体识别技术随着机器学习的发展愈加成熟，到现在已发展多种类别，例如 Ritter 等 [7] 通过开源渠道推特（Twitter）抽取网络安全领域中的开源情报，通过机器学习模型训练自动提取实体；Joshi 等 [8] 从国家漏洞数据库等开源渠道提取数据，采用条件随机场 [9]（Conditional Random Fields，CRF）进行模型训练完成命名实体识别。将这些方法汇总，可划分为无监督的学习方式、有监督的学习方式以及半监督的学习方式。

（1）基于无监督的学习方式

无监督统计方法 [10] 无需预先标注出数据

集，只需对大规模未标注语料库进行简单特征提取与分析，就能实现通用命名实体识别，最常见的方法为基于相似度的命名实体识别和基于聚类的命名实体识别。基于相似性的命名实体识别需要依赖字面、文本、同义词等进行运算，从而确定命名实体类别；基于聚类的命名实体识别一般将具有相似特征（例如词义、词性）或处于上下文的词进行聚类，而每个聚类可能代表一个命名实体类别，完成初步筛选后通过人工分析来确定每个聚类的具体含义，而人工分析依旧离不开成本和精度问题，因此无监督统计方法的命名实体识别需进一步优化。

（2）基于有监督的学习方式

为优化无监督统计方法，早期 Tran 等 [11] 提出利用隐马尔科夫模型（Hidden Markov Models，HMM）模拟人的思维来降低实体分类识别的人工费用，这种命名实体识别技术称为有监督统计方法。该方法的一般流程为利用数据标注进行特征提取，并选择合适模型进行训练，其中数据标注采用特定格式 [12-13]（如 BIO、BILUO 等）来区分命名实体的开始、内部和结束位置。有监督统计方法常用训练模型除了隐马尔科夫模型、条件随机场之外，还有支持向量机 [14]（Support Vector Machines，SVM），该方法有效解决了成本和精度问题，但新的问题接踵而至：一是有监督统计方法依赖于大量标注数据来训练模型，因此对训练数据有较高要求，开源情报虽来源广泛，但经过处理后不能保证数据的数量与质量；二是不同类型或领域数据训练效果可能差异较大，需花费大量精力对模型进行调整，针对这个问题，提出了半监督式统计方法。

（3）基于半监督的学习方式

基于半监督式统计的命名实体识别结合了有监督和无监督的优点，即将少量标注数据和大量未标注数据结合进行联合训练。半监督式统计的步骤与有监督一致，但值得注意的是，在数据标注中只标注部分数据，而未标注数据需用已通过标注数据训练好的模型进行自动标注。利用这种方式，对开源情报进行筛选与处理，选出合适的有标签和无标签数据集后训练出模型，目前 Xie 等 [15] 提出全新的概率表示对比学习框架（Probabilistic Representation Contrastive Learning，PRCL），通过引入全局分布原型和虚拟负样本形成半监督命名实体识别，有效改善无监督式鲁棒性与有监督式性能问题，为情报领域研究提供参考。三种方法的对比如表 1 所示。

表 1　命名实体识别统计方法比较

方法名称	方法特点
无监督	无需预先标注，但人工成本较高、精度较低、解释性差
有监督	准确性高、解释性强，但依赖数据质量、模型复杂度高
半监督	成本低、泛化能力强

1.1.3　基于神经网络的深度学习方法

近年来，随着机器学习的重大突破，命名实体识别相关模型也不断优化，这些优化模型下的命名实体技术采用端到端的解决思路，减少了特征工程的设计，虽对数据的数量和质量有了更高要求，但训练效果更好，训练结果更符合预期，可以说命名实体识别从大规模语料库统计跨越到一个全新的阶段——神经网络深度学习 [16] 阶段。

doi:10.3772/j.issn.2095-915x.2025.03.007

Transformer[17] 模型目前为开源情报实体链接技术的主流模型，其模型结构如图 2 所示。

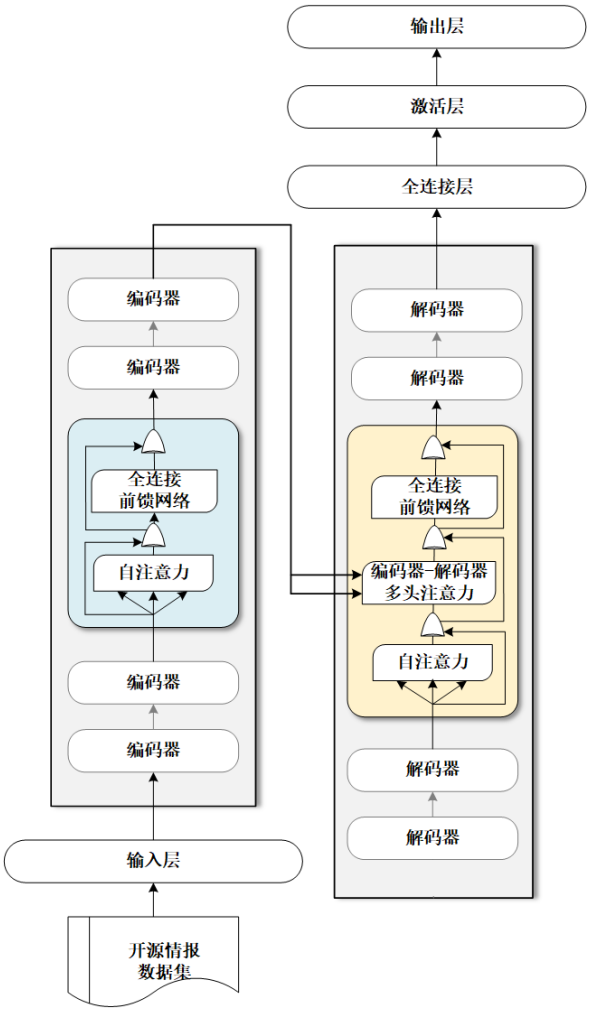

图 2　Transformer 模型结构图

从图中可明显看出 Transformer 模型包含注意力机制，且灵活运用多头注意力机制，与仅有注意力模型相比，增强对输入序列中不同部分的关注能力，提高模型在命名实体识别任务中的准确性，开源情报结构特点促使其需增强对边界与内部结构的敏感性，而 Transformer 模型中的自注意力有助于更准确地识别出命名实体的边界和类型，从而提高识别准确性。此外结合前馈网络对输入序列进行非线性变换，能捕捉出更复杂的特征。对开源情报实体链接技术来说能够更好地理解上下文中的语义信息，从而提高准确性。

随着大语言模型的不断演化，尤其是 ChatGPT 的横空出世，开源情报实体链接模型理应在此基础上进行优化，但受到权限、资金等非技术方面因素制约，开源情报实体链接模型改进效果不甚理想，但在今年 DeepSeek 的爆火有望使开源情报实体链接走向新的征程，至少在命名实体识别方面迈入新阶段。DeepSeek 仍依托于 Transformer 架构进行改进，以图 2 为参考，将多头注意力改进为多头潜在注意力（Multi-head Latent Attention，MLA），而多头潜在注意力采用低秩联合压缩技术与多头注意力机制，将键值对压缩为一个潜在向量，这样做的好处不仅仅是显著减少了缓存容量需求，优化了资源利用，提高推理效率，对于开源情报命名实体识别技术来说，还能跨越多个单词或句子，很好地捕捉文本中的长距离依赖关系，同时关注文本中不同位置和不同语义层面的信息。开源情报来源广、知识层面杂乱、结构复杂，能提升语义理解和生成能力，捕捉长距离依赖关系等同于提高命名实体识别准确率。

DeepSeek 还在 Transformer 架构的前馈网络中增加 DeepSeek 专家混合（DeepSeekMoE），即共享 [18] 和路由专家混合 [19]，其中共享专家模块始终保持激活状态，负责捕获和整合不同上下文中的通用知识，在命名实体识别任务中可以提取出文本中的通用特征，如人名、地名、机构名等常见实体的共同特征，从而帮助模型更准确地识别这些实体；路由专家模块专注于

处理特定领域的复杂问题，这种灵活性于开源情报而言能更好地适应命名实体识别任务中不同实体类别和上下文环境的多样性。当共享和路由专家模块根据不同的任务需求灵活混合时，不管是针对开源情报的命名实体识别还是实体链接，都可有效提高准确率和运行速率，因此增添 MLA 和 DeepSeekMoE 的 Transformer 架构，即 DeepSeek 蒸馏模型有望成为基于神经网络开源的情报实体链接新里程碑。

1.2 候选实体

实现命名实体识别之后，需将实体与可能的提及形式进行映射，属于实体处理相关技术。实体处理相关技术应用广泛，无论是实体链接、机器翻译还是问答系统中都需借鉴相关模型，但针对开源情报实体链接，最常见的候选实体方法还是利用词典查询。

1.2.1 基于词典的候选实体方法

在机器学习体系中，词典一般指的是由实体及其相关名称（如别名、缩写、同义词等）所组成映射关系的数据库。词典可通过多种方式构成，一种是利用现有知识库 [20]（如 Wikipedia、DBpedia、Freebase、百度百科等）进行提取，另一种是利用自然语言处理技术（NLP）进行收集，而开源情报具有领域化特点，因此还可以采用专业数据库进行提取，但无论何种方式，最后生成的词典的每个条目通常包含实体名称、唯一标识符和描述性文本。通过对词典里的实体名称与实体提及进行精细、部分或者模糊匹配，最终生成候选实体。

精细匹配，顾名思义是词典里的实体名称与实体提及完全一致才能进行匹配；部分匹配是扩大匹配范围，比如实体名称子集、相同前后缀、同义词等都可以列入候选实体；模糊匹配可将缩写、略写、扩展词都加入其中。以开源情报“在跨国救援演习中，我方 QRF（快速反应部队）迅速响应并成功营救被困人员，展现了高效处置能力”为例，通过命名实体识别方法提取“QRF”后，经过词典匹配得到的结果如表 2 所示。

表 2 基于词典的候选实体映射

方法名称	候选实体
精细匹配	QRF（快速反应部队） QRF（复苏快速反应小组）
部分匹配	QRT（快速反应小组） RTF（快速反应部队）
模糊匹配	Quick Reaction Force

1.2.2 基于统计的候选实体方法

除了利用词典，还可以利用统计方法。基于统计的候选实体方法可利用 TF-IDF[21]（Term Frequency-Inverse Document Frequency）或者 BM25[22] 生成候选实体。TF-IDF 是一种静态加权技术，因此它的计算就是简单的 TF 与 IDF 相乘，其中 *TF* 是词频，即某给定词 T 在文档 D 的词项总数 $size(D)$ 出现频率 $freq(T,D)$；*IDF* 是逆文档频率，代表文档数量 N 中给定词出现的文档次数 $df(T)$，如公式（1）所示。

$$TF\text{-}IDF = \frac{freq(T,D)}{size(D)} \times \left(-log_2 \frac{df(T)}{N} \right) \quad (1)$$

利用 TF-IDF 方法可通过得分数确定候选实体。BM25 原理与 TF-IDF 相近，只不过更关注词与句、句与文档之间的相关度，如公式（2）所示。

doi:10.3772/j.issn.2095-915x.2025.03.007

$$score(A,B)=\sum_{i=1}^{n} w_i \times R_i \quad (2)$$

其中 w_i 为 R_i 的权重，R_i 为 A 中每一个字与 B 的相关性，$score(A,B)$ 为分数。引入分数这一概念，使候选实体准确度有所提高，而之后出现的 N-gram 模型同样利用分数候选实体。

N-gram 模型 [23] 又称基于概率的判别模型，建立长度为 n 字节的窗口对开源情报滑动，统计每个词序列（gram）的出现频度，根据预先设定的阈值或规则，选择频度较高的词序列作为候选实体，即用分数候选实体。与 BM25 相比，N-gram 模型并没有采用文档长度归一化这种简单的处理方法，而是利用捕获上下文联系和依赖性提高精确性。但值得注意的是，N-gram 模型前提假定第 n 个词出现的概率只与第 n–1 个词相关，与其他词不相关，因此易产生零概率、计算效率低、距离度量失效等问题，而这些正是开源情报所具有的特点。基于此，需结合平滑技术、缓存技术或语义模型与 N-gram 模型混合使用以生成候选实体列表。

针对开源情报零概率问题，徐志明等 [24] 将 Laplace、Good-Turing、Katz 等平滑方法融合 N-gram 模型进行了经验性对比，结果表明混合平滑技术优于单 N-gram 模型。针对计算效率低问题，常采用缓存技术来减少 N-gram 模型的存储和计算成本，虽然该技术也会引发诸如缓存一致性维护和缓存淘汰策略等连锁反应，但相比开源情报特性，大部分学者更关注如何解决距离度量失效问题。

目前主流融合方式是 N-gram 模型与词嵌入（word embeddings）或上下文嵌入（context embeddings）结合，利用两者的互补性，提高模型对长距离依赖关系的捕捉能力，即解决距离度量失效。Wieting 等 [25] 将两者结合形成 Charagram embeddings，并通过非线性变换得到低维嵌入，实验结果较为理想。

1.3 候选排序

获得候选实体后，要注意并不是所有实体都能链接，还需经过候选排序来确定形成＜实体提及—候选实体＞对。

1.3.1 基于传统机器学习的候选排序方法

早期传统机器学习候选排序方法大多数采用二分类模型 [26]，即通过学习将候选实体分正反类，这种方法简单迅速，但对数据集要求极高，开源情报来源广、数量多、结构杂，因此很容易出现数据不平衡问题，除此之外，二分类模型还存在区分多个正例能力有限的问题 [27]。针对这些问题，在后期传统机器学习候选排序时采用支持向量机解决。

SVM 是一类利用超平面将最近的数据点间隔开来进行二元分类的广义线性分类器，如公式（3）所示。

$$y(x)=\varphi(x)w^t+b \quad (3)$$

SVM 最大化提高分类准确性与鲁棒性，其中公式（3）是数据集 x 的特征映射函数，w^t 是权重矩阵，b 是偏置量，通过计算得到的数值为排序得分，而得分最终决定了候选排序的先后顺序。

1.3.2 基于神经网络的候选排序方法

基于传统机器学习候选排序方法本质上依赖于知识库，当知识库出现数据稀疏问题时，大部分方法只能改为人工输入信息进行候选实体排序，基于神经网络的候选排序方

法应运而生。

基于神经网络的候选排序方法与传统机器学习候选排序方法相比，无论是原理还是方法都发生改变。首先要通过神经网络模型提取特征，这些特征可以是文本特征（词性、上下文窗口内的词汇等），可以是语义特征（预训练词向量所捕捉的语义关系），可以是上下文特征（所编码的上下文信息），也可以是实体本身具有的属性（描述、类别等），提取特征之后对标注序列进行训练，学习特征与输出之间的关联关系，在低维稠密空间为不同实体间语义关系构建同一表示，通过相似度得分排序候选实体。开源情报也不例外，近些年谢腾等 [28] 利用最新 BERT-BiLSTM-CRF 模型对 MSRA 数据集进行命名实体识别、候选实体与候选排序，并通过召回率和 F1 值证明了该方法的优越性。

2 相关改进技术

经过命名实体识别、候选实体与候选排序后，实体提及和实体概念已实现关联，通过准确性与完整性分析，发现开源情报实体链接技术虽对广泛性、多样性有一定改善，但未关注歧义性问题，换言之存在实体歧义、整体信息缺失、实体缺失与实体变形等问题。在前面已介绍实体链接技术为综合模型，起到“中转站”作用，当出现问题时，可利用自然语言相关模型进行解决。针对实体歧义问题可在命名实体识别之后增加实体消歧技术、针对整体信息缺失可通过关系抽取技术捕捉长文本所潜藏的信息进行处理、针对实体缺失与实体变形可在候选实体生成后进行处理，如图 3 所示。

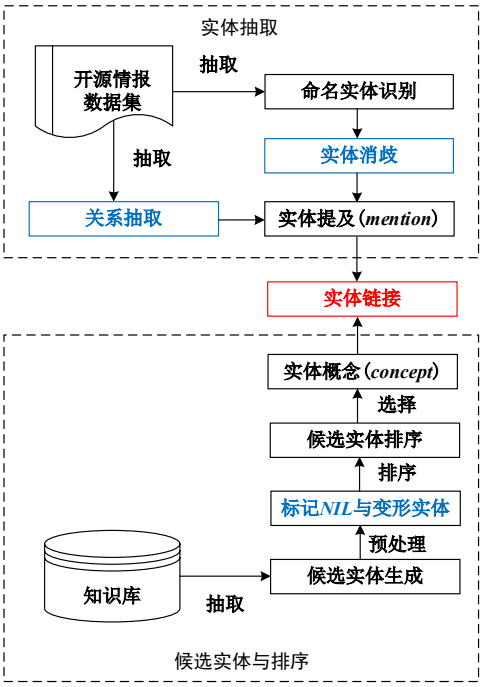

图 3　开源情报实体链接改进流程图

2.1 命名实体消歧

实体歧义指由于多个实体拥有相同命名，在实体链接时将错误的实体映射到知识库上，导致重名现象（Polysemy）降低其可靠性，而开源情报的来源广泛，即使按照领域划分也依旧出现一词多义现象，例如在军事领域中“猎豹”（Gepard）一词多指代德国某防空坦克，但在南非同样有名为“猎豹”（Rooikat）的武器系统，或可能存在其他国家或组织所命名的“猎豹”武器系统。除此之外，有些实体在名称上具有差别，例如在日常生活中“番茄”与“西红柿”虽读法不一致，但都属于同一实体，这都对实体链接技术产生影响，因此需要命名实体消歧技术 [29-31]（Named Entity Disambiguation，NED）解决这些问题。

2.1.1 实体显著性

主流命名实体消歧技术是通过实体显著性、

doi:10.3772/j.issn.2095-915x.2025.03.007

上下文相似度或语义关联度进行划分。实体显著性是最简单最直观的消歧手段，例如直接通过 Wikipedia、DBpedia、Freebase、百度百科等开源渠道中的查询量、浏览量或者点击量，按照得分进行排序，而这种方法就被称为基于流行度的实体消歧技术。

在开源情报中，基于流行度的实体消歧技术较为常见。还以“猎豹”为例，首先通过多个开源渠道（如新闻报道、社交媒体、军事论坛等）收集提及“猎豹”的所有相关信息，然后对收集到的信息进行预处理，提取实体之后统计每个“猎豹”（Gepard、Rooikat 等）在开源信息中的提及次数，对每个实体赋予权重并按照从高到低的顺序将最终得分进行排序，最终得分最高的实体优先被保留。这种消歧技术可解释性、可操作性较强，但存在致命缺陷——“冷启动”问题，随着科技日新月异，不仅在军事上新的装备与技术不断涌现，在其他领域上也不断涌出新概念新思想，就连例子中“猎豹”（Gepard）一词也早已过了德国某防空坦克专属词时代，当新“猎豹”实体指称出现时，原有的开源渠道无法提供足够数据量进行流行度的有效计算，因此该消歧技术存在一定适用局限性，甚至起到反作用。

基于表层形式相似度的实体消歧技术为避免上述问题采用了新实体显著性消歧原理，即通过编辑距离、Jaccard 相似度 [32] 计算实体名称相似度，再通过得分进行排序或设置阈值进行合并归类，比如在不同开源渠道中某型护卫舰被称为“XX 级护卫舰”“XX 型护卫舰”或“XX 号护卫舰”等，为了判断是否为同一类护卫舰，可采用该技术计算并分类，确保实体的准确性和一致性。这种方法简单有效但依旧存在问题，当出现名称完全重复或者对具有缩略词的实体进行消歧时无法得出准确结果。

2.1.2 语义关联

除了上述消歧方式，还有基于共性的实体消歧技术，但无论哪种方法，在开源领域都只适合缩小候选实体范围，而不是直接获得消歧结果，而基于语义关联的命名实体消歧技术旨在从外部知识库获取概念、属性、关系等候选实体详细信息，与实体信息进行对比，计算语义相似度。

语义消歧依赖于外部知识库，外部知识库中实体及其相关信息越详尽，消歧结果越精确。假设有一篇文章可在开源渠道上查询，其中写着这样一段话——“辽宁舰舰长约 XX 米，舰宽约 XX 米，标准排水量约 XXX 吨，载重量约 XXX 吨……作为中国第一艘服役的航空母舰实现了完全自主设计、自主建造、自主配套，体现了其庞大的规模与强大的作战能力”。通过实体显著性无法将“辽宁舰”与“中国第一艘服役的航空母舰”联系在一起，而通过外部知识库可获得“辽宁舰”和“中国第一艘服役的航空母舰”的舰体尺寸、排水量、载重量、武器或装备等指标，通过模型计算其相似度可判定“辽宁舰”和“中国第一艘服役的航空母舰”为同一实体。

2.1.3 上下文相似度

在“辽宁舰与中国第一艘服役的航空母舰”的案例中有一点值得关注，在文中不仅列出“辽宁舰”的参数指标，同时“作为”一词也间接表明了“辽宁舰”就是“中国第一艘服役的航空母舰”，这就说明同一文章、同一主题甚至同一领域的开源情报存在实体的更多线索，可以利用上下文相似度进行实体消歧。

由于开源情报实体的领域性较强，尤其是在军事上更侧重于对具体性能设备、技术等在特定情况下的实践与应用，因此常用实体线索或主题模型（Blei 等[33]提出的潜在狄利克雷分配 LDA、张伟[34]提出的非负矩阵分解 NMF 等）计算上下文相似度。如果在上下文中通过实体线索锁定候选实体实现（在“辽宁舰”和 “中国第一艘服役的航空母舰”中，具有显性的实体等价关系表达）为基于实体线索的实体消歧技术；如果根据文章上下文信息锁定主题或者计算主题相似度（ “辽宁舰”与“中国第一艘服役的航空母舰”在上下文中可知都为舰艇，可确定主题类型缩小候选实体范围，而通过文章对实体的特征描述，可计算出候选实体相关度来进行排序）为基于主题模型的实体消歧技术。无论是哪种实体消歧技术，在开源情报上其准确率和效率更高，因此基于上下文相似度的实体消歧技术更为常见，不同消歧技术对比分析如表 3 所示。

表 3　不同消歧技术对比分析

消歧技术	技术特点
实体显著性	利用流行度、表层、共性等相似度快速消歧
语义关联	依赖于外部知识库
上下文相似度	通过实体线索或主题模型计算相似度进行消歧

2.2　关系抽取

通过对基于上下文相似度的实体消歧技术的分析，可以看出在开源情报中不仅需关注单个实体与知识库之间的映射关系，同时实体之间也蕴含相应关系，正确分析有助于更精确进行实体链接，因此在进行命名实体识别时应抽取实体关系[35]。

与命名实体识别一致，关系抽取也经历三个阶段。第一阶段为基于规则的关系抽取，该方法依旧依赖于规则集，通过专家构建的语法语义或句法结构进行关系抽取，因此具有可解释性强、针对特定领域准确性高、不依赖大量数据等优点，但同样也存在覆盖率低、可扩展性差和维护成本高的问题，因此需利用统计方法解决这些问题。

第二阶段为基于统计的关系抽取，与命名实体识别的唯一区别是在关系抽取中实体与关系要同时抽取。因此在有监督学习中可通过核函数来计算不同关系实例特定表示之间的相似度，从而进行关系分类；半监督学习可使用句法模式等方法。同理，基于统计的关系抽取泛化能力强、自动化程度高，但存在领域适应性差、依赖标注数据等问题。

基于深度学习的关系抽取即为第三阶段，是目前主流的关系抽取方式，其核心在于从非结构化或半结构化的文本数据中，识别出两个或多个实体之间的特定关系，例如从开源情报“属于美国海军的核动力航空母舰‘尼米兹号’的命名源于二战时期的美国海军四星上将切斯特·威廉·尼米兹”中，可以抽取出“美国海军的核动力航空母舰‘尼米兹号’—命名源于—美国海军四星上将切斯特·威廉·尼米兹”这样的三元组关系，再进行切分又会得到“核动力航空母舰‘尼米兹号’—属于—美国海军”新三元组关系。常用的深度学习关系抽取有卷积神经网络[36]（Convolution Neural Network，CNN）、循环神经网络[37]（Recurrent Neural Network，RNN）、长短期记忆网络[38]（Long-Short Term Memory，LSTM）、注意力机制[39-40]

doi:10.3772/j.issn.2095-915x.2025.03.007

（Attention Mechanism，AM）等模型，但随着深度学习技术的发展，知识图谱[41-44]（Knowledge Graph，KG）等新技术的出现使精度和效率得到了显著提升。

知识图谱通过捕获全局信息将实体与关系按统一标准整合在同一图结构中，因此具有强扩展性与全局性，除此之外利用知识图谱进行关系抽取可以发现数据中的隐含关系和未知事实，同时知识图谱所具有的开放数据格式有利于共享和交互。若将知识图谱与最先进神经网络融合对开源情报进行关系抽取，可弥补现有缺陷，目前最热门的是图注意力机制[45]（Graph Attention Mechanism，GAT）。随着 DeepSeek 的爆火，多头注意力机制可能被多头潜在注意力机制所替代。而与知识图谱结合，是否有助于开源情报实体链接也是未来研究方向。

2.3 实体缺失与实体变形

开源情报虽具有来源广、数据多的特点，但存在实体缺失[46]、实体变形[47]等问题。

若知识库中缺失与实体提及指称相匹配的实体概念，即存在实体缺失的问题，则将该指称标记为不可链接（NIL）。这有助于识别出知识库中的空白区域，并为后续的知识库更新提供指导，例如 Dredze 等[48]使用学习排序模型对候选实体进行排名，如果候选实体利用该模型将 NIL 作为排名靠前的实体输出，则说明该实体不可链接。

若出现实体变形问题，可通过外部数据或基于深度学习的命名实体识别技术来解决，除了这两种方法之外，也可以和实体缺失一样优化候选实体排序算法，比如采用投票机制或加权求和的方式，将多种排序算法的结构进行合并。

3 未来展望

现有的开源情报实体链接改进技术已经取得了显著进步，然而随着深度学习技术的不断突破，面向开源情报的实体链接技术只关注实体准确性与完整性方面改进，远不能满足日益增长的多元化和个性化需求，DeepSeek 的出现有望为未来开源数据实体链接改进提供新范式。

3.1 多源知识融合实体链接技术

现阶段针对开源情报层面的实体链接技术主要解决异构性问题，而多样性却较为忽略，最典型的案例就是大众对开源情报的感知只停留在文字方面，而对图像、声音、传感器等开源数据挖掘具有重要价值。如何对不同来源、不同模态的开源情报进行融合，通过实体链接形成更完善、更准确的知识体系将是未来研究热点。目前多源知识融合实体链接技术[49-52]侧重于改进深度学习模型，如领域多模态知识图谱、针对实体链接的变体 BERT 模型[53-54]——Liu 等[55]在原有 BERT 模型基础之上增加动态掩码和注意力头数所诞生的新模型 RoBERTa、Lan 等[56]引入跨层参数共享机制后诞生新模型 ALBERT 等，这些模型都在一定程度上得到较好结果，但都没有达到预期目标。DeepSeek 中 Janus-Pro-7B 多模态模型采用新模型架构，打破固有改进思维，在未来有望成为新的发展方向。

3.2 跨语言实体链接技术

跨语言实体链接技术[57]旨在解决不同语言之间实体链接的问题，通过处理语言的差异性、充分利用多语言之间共性和互补性来实现跨语言实体链接。开源情报来源广泛，不拘泥于同一地

区，因此出现不同语言问题也不足为奇，但在自然语言处理中如何将跨语言准确无误归一化进行处理不仅仅是机器翻译的痛点，还是实体链接的痛点，究其原因在于无合适方式、无足够量训练数据（大部分训练数据为英语语种）使同种领域迁移效果未达到标准，因此目前英文实体链接技术较为成熟，中文等其他语种实体链接技术仍处于发展阶段。DeepSeek 采用层次化语言建模框架，包含专门的语言适配层，可动态调整语法规则与语义表征，并利用跨语言知识迁移，将中英语言知识转化为目标语种的语义表征，实现快速扩展支持新语种，从而降低多语言部署成本，未来可成跨语言主流研究方向。

3.3 小样本、零样本实体链接技术

开源情报虽然获取渠道充足，但经过数据处理后，某些特定领域受到同质化、时效性或数据投毒等问题影响，因此用于实体链接的训练样本不足以充分支撑学习实体之间的复杂语义关系，即使利用外部多源知识库进行匹配，由于不同知识库对实体描述可能存在差异，导致知识不匹配的情况产生，因此需要小样本甚至零样本实体链接技术。目前 Li 等 [58] 提出的元学习（DeepSeek 也采用该方法）、Snell 等 [59] 提出的几何原型等网络模型使小样本实体链接得到突破，而针对此提出的零样本实体链接技术 [60-61]（Zero-shot Entity Linking，ZEL）还在研究初级阶段，在未来可能会有所突破。

3.4 动态实体链接技术

除了从开源情报本身特性出发，还可从其研究价值入手探究实体链接技术未来发展方向。实时开源情报可用以支持决策制定、风险评估、安全监测等活动，因此其研究价值在于如何在竞争激烈的环境中获得优势。为满足实时数据处理的需求，动态实体链接技术 [62] 需在极短时间内实现实体识别与实体链接，虽有相关研究支撑动态实体链接技术，但对“极短时间”的判定仍不能满足大众需求，DeepSeek 中 MoE 架构实现了模型容量的高效扩展。这种架构使得模型能够根据不同的输入任务动态地选择合适的专家进行处理，从而提高了模型的泛化能力和鲁棒性，为动态实体链接技术指明发展道路。

4 结语

本文通过整理近十年开源情报及人工智能领域论文，将开源情报下实体链接及关联技术进行系统综述研究，构建其知识脉络主线。在梳理过程中，不仅回顾该技术理论基础与发展历程，结合最新人工智能领域产物 DeepSeek 研究其在开源情报实体链接技术的关联，同时还探讨未来可能面临的挑战与技术创新点。除了多模态、小样本等发展方向，还有数据隐私下保护可信、自动错误纠正机制等方向仍需完善，因此面向开源情报的实体链接技术研究任重道远，需在后续的工作中不断挖掘与改进。

参 考 文 献

[1] 范昊 , 郑小川 . 国内外开源情报研究综述 [J]. 情报理论与实践 , 2021, 44(10): 185-192, 201.
[2] 冯兴杰 , 彭洲 , 张成豪 , 等 . 民航领域突发事件的实体链接方法 [J]. 计算机应用研究 , 2023, 40(4): 1052-1058, 1064.
[3] 刘航冶 , 富铁楠 , 杨勇 . 互联网开源文本情报智能分析技术综述 [J]. 情报杂志 , 2023, 42(2): 12-16.
[4] SHEN W, WANG J, HAN J. Entity linking with a

doi:10.3772/j.issn.2095-915x.2025.03.007

knowledge base: issues, techniques, and solutions[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 27(2): 443-460.
[5] NADEAU D, SEKINE S. A survey of named entity recognition and classification[J]. Lingvisticae Investigationes, 2007, 30(1): 3-26.
[6] 孙青 . GATE：一个开源的自然语言处理平台 [C]// 2006 年“数字图书馆与开放源码软件 (OSS)”学术研讨会 , 2006: 88-94.
[7] RITTER A, CLARK S, ETZIONI O. Named entity recognition in tweets: an experimental study[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 1524-1534.
[8] JOSHI P, DING R, LIU Y. Extracting cybersecurity entities from natural language text using CRFs[C]// Proceedings of the 2017 IEEE International Conference on Semantic Computing (ICSC). IEEE, 2017: 254-257.
[9] CHEN W, ZHANG Y, ISAHARA H. Chinese named entity recognition with conditional random fields[C]// SIGHAN 2006, Sydney, Australia. 2006: 118-121.
[10] LUO Y, ZHAO H, ZHAN J. Named entity recognition only from word embeddings[J]. arXiv: 1909. 00164, 2020.
[11] TRAN K, BISK Y, VASWANI A, et al. Unsupervised neural hidden markov models[J]. arXiv: 1609.09007, 2016.
[12] 高劲松 , 周习曼 , 梁艳琪 . 面向关联数据的实体链接发现方法研究 [J]. 中国图书馆学报 , 2016, 42(6): 85-101.
[13] 杜雪盈 , 刘名威 , 沈立炜 , 等 . 面向链接预测的知识图谱表示学习方法综述 [J]. 软件学报 , 2024, 35(1): 87-117.
[14] LEE K J, HWANG Y S, KIM S, et al. Biomedical named entity recognition using two-phase model based on SVMs[J]. Journal of Biomedical Informatics, 2004, 37(6): 436-447.
[15] XIE H, WANG C, ZHAO J, et al. PRCL: probabilistic representation contrastive learning for semi-supervised semantic segmentation[J]. arXiv: 2402. 18117, 2024.
[16] BOZIC V, DORDEVIC D, COPPOLA D, et al. Rethinking attention: exploring shallow feed-forward neural networks as an alternative to attention layers in transformers. Association for the Advancement of Artificial Intelligence[J]. arXiv: 2311. 10642, 2024.
[17] CHOWDHURY M N R, ZHANG S, WANG M, et al. Patch-level routing in mixture-of-experts is provably sample-efficient for convolutional neural networks[C]// Proceedings of the 40th International Conference on Machine Learning (ICML). ACM, 2023: 6074-6114.
[18] FEDUS W, ZOPH B, SHAZEER N. Switch transformers: scaling to trillion parameter models with simple and efficient sparsity[J]. arXiv: 2101. 03961, 2021.
[19] LI J, SUN A, HAN J, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(1): 50-70.
[20] 曾宇涛 , 林谢雄 , 靳小龙 , 等 . 基于多维信息融合的知识库问答实体链接 [J]. 模式识别与人工智能 , 2019, 32(7): 642-651.
[21] 牛萍 . TF-IDF 与规则结合的中文关键词自动抽取研究 [D]. 大连：大连理工大学 , 2015.
[22] 陈乐乐 , 黄松 , 孙金磊 , 等 . 基于 BM25 算法的问题报告质量检测方法 [J]. 清华大学学报（自然科学版）, 2020, 60(10): 829-836.
[23] BROWN P F, DELLA P V J, DE SOUZA P V, et al. Class-based n-gram models of natural language[J]. computational linguistics, 1992, 18(4): 467-480.
[24] 徐志明 , 王晓龙 . N-gram 语言模型的数据平滑技术 [J]. 计算机应用研究 , 1999, 16(7): 21-24.
[25] WIETING J, BANSAL M, GIMPEL K, et al. Charagram: embedding words and sentences via character n-grams[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, CA, USA, 2016, 1556-1567.
[26] SHEN W, LI Y, LIU Y, et al. Entity linking meets deep learning: techniques and solutions[J]. arXiv: 2109. 12520, 2021.
[27] 李智杰 , 王瑞 , 李昌华 , 等 . 联合关系上下文负采样的知识图谱嵌入 [J]. 数据分析与知识发现 , 2022, 6(12): 90-98.
[28] 谢腾 , 杨俊安 , 刘辉 . 基于 BERT-BiLSTM-CRF 模型的中文实体识别 [J]. 计算机系统应用 , 2020, 29(7): 48-55.
[29] 温萍梅 , 叶志炜 , 丁文健 , 等 . 命名实体消歧研究进展综述 [J]. 数据分析与知识发现 , 2020, 4(9): 15-25.
[30] 何儒汉 , 唐娇 , 史爱武 , 等 . 基于实体消岐和多粒度注意力的知识库问答 [J]. 计算机工程与设计 , 2022, 43(2): 560-566.
[31] 龚凌晖 . 中文命名实体识别与歧义消解研究 [D]. 上海：复旦大学 , 2011.
[32] 俞婷婷 , 徐彭娜 , 江育娥 , 等 . 基于改进的 Jaccard 系数文档相似度计算方法 [J]. 计算机系统应用 , 2017, 26(12): 137-142.
[33] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet

allocation[J]. Machine Learning Research Archive, 2003(3): 993-1022.
[34] 张伟 . 非负矩阵分解新方法研究 [D]. 西安：西安电子科技大学 , 2013.
[35] 赵国清 , 何佳洲 , 乔慧 , 等 . 基于远程监督关系抽取的开源情报处理 [J]. 指挥控制与仿真 , 2021, 43(1): 69-73.
[36] 姚建军 , 李剑宇 , 岳昆 , 等 . 基于概率推理的知识图谱链接预测方法 [J]. 计算机集成制造系统 , 2023, 29(10): 3483-3495.
[37] 刘华玲 , 孙毅 . 基于实体识别和信息融合的知识图谱研究——以新冠肺炎疫情为例 [J]. 计算机技术与发展 , 2022, 32(9): 107-113.
[38] 郭世伟 , 马博 , 马玉鹏 , 等 . 基于预训练模型和图卷积网络的中文短文本实体链接 [J]. 中文信息学报 , 2022, 36(12): 104-114.
[39] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[C]// Proceedings of the International Conference on Learning Representations (ICLR). OpenReview, 2018: 1-15.
[40] 吴晓崇 , 段跃兴 , 张月琴 , 等 . 基于 CNN 和深层语义匹配的中文实体链接模型 [J]. 计算机工程与科学 , 2020, 42(8): 1514-1520.
[41] GEORGIADOU D, DIAKOLOUKAS V, TSIARAS V, et al. Clock work-RNN based architectures for slot filling[C]// Interspeech 2017, Stockholm, Sweden. 2017: 2481-2485.
[42] PENG Z, WEI S, JUN T, et al. Bidirectional long short-term memory networks for relation classification[C]// Proc. 29th Pacific Asia Conf. Lang., Inf. Comput. (PACLIC), Shanghai, China, 2015: 73-78.
[43] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need. Advances in neural information processing systems[C]// Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NeurIPS), Long Beach, CA, USA, 2017: 5998-6008.
[44] 付林 , 刘钊 , 邱晨 , 等 . 基于特征联合与多注意力的实体关系链接 [J]. 计算机工程 , 2022, 48(8): 53-61.
[45] 王晓狄 , 黄诚 , 刘嘉勇 . 面向网络安全开源情报的知识图谱研究综述 [J]. 信息网络安全 , 2023, 23(6): 11-21.
[46] 李圣杰 , 周新 , 史一民 . 基于协同注意力和递归随机游走的实体链接方法 [J]. 计算机应用与软件 , 2022, 39(8): 163-170, 186.
[47] 封皓君 , 段立 , 张碧莹 , 等 . 双向循环进化的实体链接及知识推理框架 [J]. 系统工程与电子技术 , 2022, 44(9): 2878-2885.
[48] DREDZE M, MCNAMEE P. Learning to rank for named entity recognition[C]// Proceedings of the ACL 2010 Conference Short Papers. Association for Computational Linguistics, 2010: 257-262.
[49] SIL A, KUNDU G, FLORIAN R, et al. Neural cross-lingual entity linking[J]. arXiv: 1712. 01813, 2017.
[50] 蒋雪瑶 , 力维辰 , 刘井平 , 等 . 基于多模态模式迁移的知识图谱实体配图 [J]. 计算机工程 , 2022, 48(8): 70-76.
[51] 孙泽群 , 崔员宁 , 胡伟 . 基于链接实体回放的多源知识图谱终身表示学习 [J]. 软件学报 , 2023, 34(10): 4501-4517.
[52] 陈烨 , 周刚 , 章梦礼 , 等 . 基于图像识别的跨模态实体链接模型 [J]. 计算机应用研究 , 2022, 39(5): 1422-1426, 1431.
[53] 蒲秋梅 , 黄方俐 , 王辉 . 基于 BERT-LSTM 模型的情感分析研究 [J]. 中国电子科学研究院学报 , 2023, 18(10): 912-920.
[54] 詹飞 , 朱艳辉 , 梁文桐 , 等 . 基于 BERT 和 TextRank 关键词提取的实体链接方法 [J]. 湖南工业大学学报 , 2020, 34(4): 63-70.
[55] LIU Y, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[J]. arXiv:1907. 11692, 2019.
[56] LAN Z, CHEN M, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[J]. arXiv:1909. 11942, 2019.
[57] 陈阳 , 万卫兵 . 多通道特征融合的实体链接模型泛化性能优化 [J]. 计算机工程与应用 , 2023, 59(16): 125-134.
[58] LI J, CHIU B, FENG S, et al. Few-shot named entity recognition via meta-learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(9): 4245-4256.
[59] SNELL J, SWERSKY K, ZEMEL R S. Prototypical networks for few-shot learning[J]. arXiv: 1703. 05175, 2017.
[60] LOGESWARAN L, CHANG M W, LEE K, et al. Zero-shot entity linking by reading entity descriptions[C]// ACL 2019, Florence, Italy. 2019: 3449-3460.
[61] ALY R, VLACHOS A, MCDONALD R. Leveraging type descriptions for zero-shot named entity recognition and classification[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021), 2021: 1516-1528.
[62] HOU F, WANG R, HE J, et al. Improving entity linking through semantic reinforced entity embeddings[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020). 2020: 6843-6853.