Computer Science and Application 计算机科学与应用, 2025, 15(5), 564-572
Published Online May 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.155129

Hans汉斯

基于Longformer和对比学习的实体链接方法

赵占芳，刘佳琪

河北地质大学信息工程学院，河北 石家庄

收稿日期：2025年4月7日；录用日期：2025年5月9日；发布日期：2025年5月21日

摘　要

实体链接是自然语言处理中的核心任务，旨在将文本中的实体与知识库中的标准实体进行精确匹配。为了应对传统的实体链接方法在长文本处理时难以捕捉长距离依赖关系，本文提出了一种基于Longformer和对比学习的端到端实体链接方法，用更加精细化的模块设计来专门优化长文本中的实体链接性能。通过融合Longformer在长序列文本处理中的优势与对比学习在优化实体匹配过程中的机制，设计了一个联合优化实体识别和链接任务的框架，从而有效提升了模型在复杂文本中的表现。在实验验证中，本文采用了AIDA-CoNLL英文标准数据集，结果表明所提出的方法在F1值上相较于现有主流方法有了1.8%至11.8%的显著提升。

关键词

实体链接，知识图谱，预训练模型，对比学习

Entity Linking Method Based on Longformer and Comparative Learning

Zhanfang Zhao, Jiaqi Liu

College of Information Engineering, Hebei GEO University, Shijiazhuang Hebei

Received: Apr. 7th, 2025; accepted: May 9th, 2025; published: May 21st, 2025

Abstract

Entity linking is a core task in natural language processing, aiming to precisely match entities mentioned in text with their corresponding standardized entities in a knowledge base. To address the challenges faced by traditional entity linking methods in capturing long-range dependencies within lengthy texts, this paper proposes an end-to-end entity linking approach based on Longformer and contrastive learning. The method incorporates a more refined module design specifically tailored to optimize entity linking performance in long texts. By leveraging the strengths of Longformer in

文章引用: 赵占芳, 刘佳琪. 基于 Longformer 和对比学习的实体链接方法[J]. 计算机科学与应用, 2025, 15(5): 564-572.
DOI: 10.12677/csa.2025.155129

processing long-sequence texts and the mechanisms of contrastive learning for enhancing entity matching, we design a unified framework that jointly optimizes entity recognition and linking tasks, thereby significantly improving the model's performance on complex texts. In the experimental evaluation, we utilize the AIDA-CoNLL benchmark dataset, and the results demonstrate that the proposed method achieves a notable improvement in F1 score, ranging from 1.8% to 11.8%, compared to existing state-of-the-art methods.

Keywords

Entity Linking, Knowledge Graph, Pre-Trained Mode, Contrastive Learning

1. 引言

实体链接旨在将文本中提到的实体指称项链接到知识库中对应的实体，是自然语言处理领域的一项重要任务[1]。实体链接的通用流程如下图 1 所示，输入文本经过预处理后进入文本编码器，再经过实体检测得到文本中存在的实体、实体消岐在候选集中选择最符合的实体。实体链接在信息抽取、问答系统、知识图谱构建等应用中发挥着至关重要的作用。在信息抽取中，实体链接可以帮助识别文本中的关键实体并将其与知识库中的结构化信息关联；在问答系统中，实体链接能够准确理解用户问题中的实体指称项，从而提供更精确的答案；在知识图谱构建中，实体链接是将非结构化文本中的实体与知识图谱中的节点对齐的关键步骤[2]。

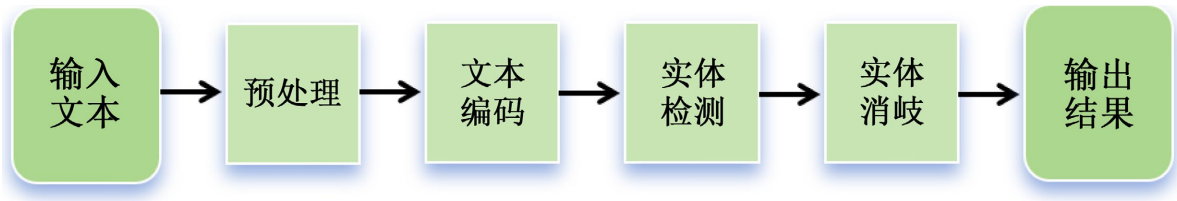

Figure 1. General flowchart of entity linking
图 1. 实体链接通用流程图

然而，实体链接面临着诸多挑战如下，首先，指称项歧义性是实体链接的核心难题之一[3]。同一个实体指称项可能对应知识库中多个不同实体。这种歧义性要求模型能够根据上下文语境准确推断指称项的真实含义。其次，上下文依赖性进一步增加了实体链接的复杂性。指称项所指代的实体往往依赖于上下文语境，例如“他去了北京”中的“北京”指的是城市，而“北京提出了新的政策”中的“北京”则指的是政府机构。此外，许多实际应用场景涉及长文本，例如新闻文章、科技文献等，如何有效地建模长文本中的语义信息并捕捉指称项与实体的关联关系是一个难题。长文本中的指称项可能分布在不同的段落或章节中，传统的模型由于输入长度限制，难以捕捉长距离依赖关系。

近年来，深度学习技术在实体链接任务上取得了显著进展。随着预训练语言模型的发展，例如 BERT、RoBERTa 等在大规模语料库上进行预训练，能够学习到丰富的语义表示，为实体链接任务提供了强大的基础。这些模型通过捕捉上下文信息，显著提升了指称项消歧和实体链接的准确性[4]。然而，传统的预

训练语言模型通常对输入长度有限制(例如 BERT 的最大输入长度为 512 个 token)，难以直接应用于长文本实体链接任务。此外，现有的实体链接方法大多依赖于标注数据进行训练，而标注数据的获取成本高昂，限制了模型的泛化能力。特别是在低资源领域或新兴领域，标注数据的缺乏使得模型的性能难以保证。

为了解决上述问题，我们的研究专注于改善实体链接模型在复杂语境和长文本中的限制，本文提出了一种基于 Longformer 预训练语言模型[5]和对比学习的实体链接创新性模型，在保证准确性的同时提高计算效率。本文所做出的主要贡献可归纳为以下三点：

1. 引入 Longformer 预训练语言模型。Longformer 是一种能够处理长文本的预训练语言模型，它通过引入滑动窗口注意力机制，能够有效地捕捉长文本中的语义信息。本文利用 Longformer 对文本进行编码，并计算指称项与候选实体之间的语义相似度。与传统模型相比，Longformer 能够处理更长的输入序列，从而更好地建模长文本中的指称项与实体之间的关联关系。

2. 引入对比学习策略对文本嵌入进行监督优化。对比学习的核心是通过比较正样本(正确的实体链接)和负样本(错误的实体链接)，让模型学会区分正确的实体链接和不正确的实体链接。通过这种方式，模型可以学习到正确数据的语义结构，能够增强模型对噪声和错误链接的抵抗能力。

3. 实验验证与性能提升。在英文 AIDA-CoNLL 数据集上的实验表明，该模型在 F1 分值上取得了与现有主流模型相比较好的结果，可以实现更高的链接精度，本文方法在处理长文本和复杂语境下的实体链接任务中具有显著优势。

2. 相关工作

实体链接任务的研究主要分为两类：基于符号逻辑的方法和基于统计学习的方法[6]。早期的工作主要集中在基于符号逻辑的方法上。这类方法依赖于人工定义的规则、模板和词典等符号化表示，通过逻辑推理实现实体链接。例如，研究者通过构建领域特定的规则库，利用指称项的表面形式(如字符串匹配)和上下文特征(如关键词、句法结构)来推断其对应的知识库实体。此外，一些方法还结合了词典资源(如 Wikipedia、Freebase)来扩展指称项的候选实体集合，并通过手工设计的特征进行实体消歧。尽管基于符号逻辑的方法在某些特定领域中表现出较高的准确率，但其局限性也非常明显[7]。这类方法的泛化能力较差，难以处理复杂的语言现象，且它们严重依赖大量的人工特征工程，需要领域专家设计规则和模板，成本高昂且难以扩展。由于符号逻辑方法的灵活性和适应性有限，难以应对开放领域和动态变化的场景需求。随着时代发展，数据规模和复杂性都在提升，基于符号逻辑的方法逐渐被基于统计学习的方法所取代。

基于统计学习的方法可以进一步分为传统机器学习方法和深度学习方法[8]。传统机器学习方法(如支持向量机、条件随机场)通常依赖于手工设计的特征(如词袋模型、TF-IDF、上下文词向量)来训练分类器或排序模型。尽管这些方法在一定程度上缓解了符号逻辑方法的局限性，但其性能仍然受限于特征工程的质量和规模。在 2018 年之前实体链接方法通常分为实体检测和实体消岐这两个步骤，Nikolaos Kolitsas 等人[9]提出了第一个神经端到端实体链接模型，并展示了联合优化实体识别和链接的好处，并且证明了工程特征几乎可以完全被现代神经网络取代。深度学习方法能够从大规模语料库中自动学习实体和指称项的分布式表示[10]，并通过计算语义相似度实现实体链接，还克服了传统方法在特征工程方面的局限性，实现了从数据中自动提取有效特征的突破[11]。基于神经网络的模型，如卷积神经网络、循环神经网络等能够捕捉指称项与上下文之间的复杂语义关系。

为了提升模型对文本的处理能力，一些研究尝试将预训练语言模型应用于实体链接任务[12]。预训练的语言模型在实体链接任务中发挥关键作用，提供了丰富的上下文和实体表示。这些模型在学习上下文表示时能够捕捉丰富的语义信息，包括实体之间的关系和语境中的重要信息[13]。与其它 NLP 任务相似，

最近的实体链接模型通常使用基于 transformer 的预训练表示方法。近年来，研究者们广泛采用增强型 BERT 及其衍生模型来提升实体链接任务的性能[14]。其中 Peters 等人[15]则探索了将多源知识库信息整合到 BERT 模型的深层结构中。在实体知识表示方面，Samuel Broscheit [16]的研究证实，在 BERT 框架中引入额外的实体知识学习机制能够显著提升实体链接效果。实体链接任务的核心挑战在于如何准确地将文本中的实体指称与知识库中的实体进行匹配，高效处理长文本并降低时间和资源消耗仍是当前研究的重难点[17]。本文提出的方法结合了 Longformer 和对比学习的端到端实体链接模型优点，能够有效地处理长文本实体链接任务，结合文本全局信息并提升模型的泛化能力。

3. 模型

本文提出的端到端实体链接模型主要由 3 个基本模块组成，分别为：文本嵌入模块、实体检测模块、链接消岐模块。总体架构如图 2 所示。Longformer 预训练语言对输入序列进行编码得到输入文本向量。实体检测模块由两个前馈神经网络组成的分类器，用于预测实体起始位置概率和结束位置概率。实体链接模块使用 LSTM 生成候选实体的文本标识符，然后使用分类器对候选实体进行重新排序。

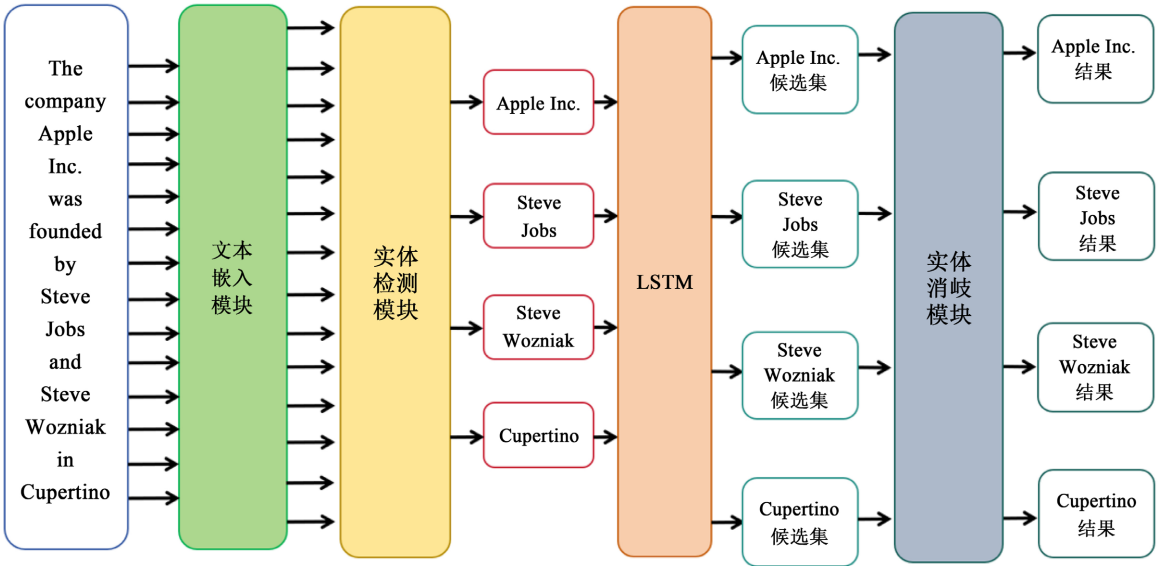

Figure 2. Model architecture diagram
图 2. 模型结构图

3.1. 文本嵌入模块

文本嵌入模块的主要任务是将输入文本转换为高维向量表示，为后续实体链接过程提供基础表征信息[18]。针对本研究场景中文本长度较长、实体上下文依赖强等特点，本文选用 Longformer 作为文本嵌入模型。Longformer 是一种改进的 Transformer 模型，专为长文本处理设计。其核心创新点在于提出了滑动窗口注意力机制，结合全局注意力机制，使模型能够高效捕捉远距离依赖信息，从而提升长文本表示能力。这种设计有效解决了传统 Transformer 架构在长文本处理中的计算复杂度过高、远距离信息捕捉能力不足等问题，特别适合文本长度较大的实体链接任务。

Longformer 的基本架构延续了 BERT 的 Transformer Encoder 框架，其预训练模型共有 12 个编码层，每个编码层包含 12 个自注意力头，每个注意力头负责捕捉不同尺度的语义信息。每个编码层的隐藏层维度为 768，这种高维表示能够充分保留文本的细粒度语义特征。考虑到实体链接任务对语义表示的实时

性要求，以及模型计算效率的权衡，本文并未使用完整的 12 层预训练模型，而是选取了 Longformer 的前 10 层作为文本嵌入模块的核心部分。研究表明，在诸多自然语言理解任务中，Longformer 的前 10 层已具备较强的语义表达能力，且计算开销相对可控。这种裁剪策略不仅降低了模型的时间复杂度和显存占用，还在保证表示能力的前提下提高了训练与推理效率。

此外，为进一步提升嵌入表示的有效性，本文采用了对比学习策略对文本嵌入进行监督优化。通过构建正负样本对，推动模型区分真实候选实体与干扰实体，从而提升嵌入表征的判别力。对于每个实体提及，其在知识库中正确的实体，即人工标注的实体链接结果，被视为正样本。从候选实体集中随机选择若干个不属于正样本的实体，作为负样本。为了增强对比学习的效果，我们确保每个提及的负样本数量与正样本数量相等，这样可以使模型的训练更加平衡，避免类别不均衡带来的偏差。在构造负样本时，本实验采用随机抽样策略，从候选实体集中随机挑选若干个非正确实体，保证负样本的多样性。

文本嵌入模块生成的高维向量不仅承载了局部上下文中的语义信息，还通过 Longformer 的全局注意力机制融入了跨句级别的长距离关联，为后续的实体候选集排序和消歧提供了更充分的语义依据。

3.2. 实体检测模块

在本研究提出的实体链接框架中，实体检测是实现端到端链接任务的关键环节之一。我们设计了一种基于 Longformer 嵌入，再与分类器结合的双位置预测机制，对实体的起始位置和结束位置分别进行预测。这种方式不仅保留了上下文信息，还能有效捕捉实体边界特征，提升检测精度。

具体而言，实体检测模块首先接收 Longformer 编码器输出的隐藏状态表示，利用起始位置分类器预测每个 Token 作为实体起始位置的概率分布。起始位置分类器主要由归一化层、全连接层、激活函数以及 Dropout 等组成。分类器首先对隐藏状态进行 LayerNorm 归一化处理，并通过 Dropout 增强鲁棒性。随后，归一化后的表示输入至线性层，将高维的 Longformer 隐藏向量(768 维)压缩至 128 维的特征表示，再经过 ReLU 激活函数引入非线性建模能力。接着，重新归一化并再次 Dropout，最后通过一个全连接层预测出每个 Token 作为起始位置的概率值。通过移除最后的预测维度，可以获得最终的起始位置概率分布。

对于实体的结束位置预测，则需要同时考虑上下文信息以及实体的起始位置先验信息。因此，结束位置分类器不仅接收批量文本对应的 Longformer 隐藏状态，还会额外接收起始位置的偏置信息。具体而言，我们对隐藏状态表示进行填充操作，将起始位置嵌入到相应位置后，与上下文表示拼接形成新的特征表示，输入至结束位置分类器中。结束位置分类器的网络层次与起始位置分类器相似，其网络层次与起始位置分类器相似，但输入维度由单一的 768 维扩展为 768×2 维，以容纳起始位置特征与上下文特征的融合信息。这种结构设计有助于模型充分利用先验信息，引导结束位置预测更加精准，特别是对于长实体和嵌套实体等复杂情况尤为有效。

对实体的起始位置进行预测时使用二元交叉熵损失函数。对结束位置进行预测时使用交叉熵损失函数。这两个损失函数的计算涉及了真实标签和模型预测的比较，以及一些额外的权重和掩码的处理，计算公式如下：

$$L_{\text{start}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\left(\hat{y}_i\right)+\left(1-y_i\right)\log\left(1-\hat{y}_i\right)\right] \tag{1}$$

$$L_{\text{end}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{i,k}\log\left(\hat{y}_{i,k}\right) \tag{2}$$

其中 N 是样本数量，是是样本 i 的真实标签的第 k 个类别的概率，是模型对样本 i 的预测输出的第 k 个类别的概率。

综上所述，起始位置和结束位置分类器协同作用，共同完成实体边界识别任务，极大提升了实体检

测的准确率和鲁棒性，为后续的候选实体生成与对比学习提供了高质量的候选集基础。

3.3. 链接消岐模块

为了有效完成候选实体的排序与最终消歧判断，本文设计了一种结合单向长短时记忆网络 LSTM 和多层感知机 MLP 分类器的实体链接模块。该模块充分考虑了长文本实体链接过程中存在的多义性问题，同时为降低计算复杂度，仅采用单向 LSTM 结构，在保证性能的前提下显著提升了计算效率。

首先，基于 Longformer 编码器获得的文本序列嵌入表示，依次传递至单向 LSTM 网络，进一步捕捉序列内部的上下文依赖关系。LSTM 网络具有较强的序列建模能力，能够对远距离上下文信息进行聚合，同时有效保留先前时序信息，从而为实体链接提供更加丰富的表征信息。假设 Longformer 输出的文本序列表示为：

$$H=\left[h_1,h_2,\cdots,h_L\right]\in R^{L\times 768} \tag{3}$$

则 LSTM 网络的隐藏状态计算过程为：

$$\tilde{h}_t=\mathrm{LSTM}\left(h_t,\tilde{h}_{t-1}\right) \tag{4}$$

其中 h_t 为第 t 个位置的上下文融合表示，进一步送入分类器进行最终的候选实体排序与实体链接判断。

$$x=\mathrm{LayerNorm}\left(\tilde{h}_t\right) \tag{5}$$

$$x=\mathrm{Dropout}\left(x\right) \tag{6}$$

$$x=\mathrm{ReLU}\left(W_1\cdot x+b_1\right) \tag{7}$$

$$x=\mathrm{LayerNorm}\left(x\right) \tag{8}$$

$$x=\mathrm{Dropout}\left(x\right) \tag{9}$$

$$p=\mathrm{Softmax}\left(W_2\cdot x+b_2\right) \tag{10}$$

其中，p 表示候选实体的类别分布(正确候选为正类，其他候选为负类)，W_1,W_2 分别为线性层参数。损失函数方面，该模块采用标准的交叉熵损失函数：

$$\mathcal{L}_{\text{link}}=-\sum_c y_c\log p_c \tag{11}$$

其中 y_c 为候选实体的真实标签，p_c 为预测的类别概率。

4. 实验

4.1. 数据集及处理

本研究采用的标准英语 AIDA-CoNLL 数据集[19]是目前规模最大且人工标注的实体链接数据集之一，数据来源于路透社新闻。该数据集被划分为 AIDA-train、AIDA-A 和 AIDA-B 三部分，总计包含的文档数为 1388 篇，总计提及数为 27816 次，平均每篇文档包含约 20 次提及实体。根据每篇文档中实体提及的数量，可以推断该数据集中的文档属于较长的文本序列。具体的统计数据请参见表 1。

Table 1. Statistics of the AIDA-CoNLL standard splits dataset
表 1. AIDA-CoNLL 标准拆分数据集的统计信息

AIDA-CoNLL	文档数	提及数
AIDA-train (训练集)	942	18540

续表

AIDA-A (验证集)	216	4971
AIDA-B (测试集)	230	4485

数据处理步骤如下，先加载 AIDA-CoNLL 数据集，每个数据样本包含输入文本、实体提及的锚定位置(anchors)和可能的实体链接候选项(candidates)，对输入文本进行分词、截断和填充，以适应模型的输入要求。将处理后的数据以 PyTorch 张量的形式返回，适用于训练模型。

4.2. 评测标准

在实体链接的评估过程中，常用的评价指标有精确率、召回率以及 F1 值。精确率是指在结果标识的实体中，有多少比例是准确的实体。召回率则是反映了模型在识别知识库中真实实体方面的能力，知识库中真实存在的实体被模型正确识别的比例。Micro-F1 是评估实体链接系统性能的重要指标之一，它综合考虑了准确率和召回率，能够有效地反映模型的整体性能[20]，本文采用 Micro-F1 来做评价指标，其具体计算公式如下：

$$P = \frac{TP}{TP + FP} \times 100\% \tag{12}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{13}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \times 100\% \tag{14}$$

其中：P 为精确率、R 为召回率；TP 是系统将文本中的实体提及正确链接到了知识库中相应的实体数量；FP 是系统将文本中的实体提及错误链接到了知识库中相应的实体数量；FN 是系统未能将文本中的实体提及链接到了知识库中相应的实体进行的数量。

4.3. 实验配置

本文实验使用的显卡是 NVIDIA GeForce RTX 3090，Python 版本是 3.9，PyTorch 版本是 2.0.1，pytorch-lightning 版本是 1.3.0。其它具体实验参数配置见下表 2。

Table 2. Experimental parameter configuration
表 2. 实验参数配置

参数名称	值
batch size	16
epoch	70
Longformer 层数	10
Longformer 学习率	2×10^{-4}
其它学习率	1×10^{-3}
优化算法	Adam
最大输入长度	4096
dropout	0.1
LSTM 的 hidden size	768

4.4. 对比实验结果与分析

为了验证本研究提出模型的有效性，设置了对比实验。与近几年实体链接模型在 AIDA-CoNLL 数据

集上的对比结果如下表 3。BiLSTM + long range context attention 方法是提出的第一个神经端到端实体链接系统，但该方法考虑所有可能的跨度作为潜在的提及，导致很高的时间复杂度和内存消耗。KnowBert 方法将 WordNet 和 Wikipedia 的子集集成到 BERT 中，以获得知识增强的 BERT。为了提高数据准备效率，该数据集偏向于较短的序列，导致其在长文本上的表现较差。BERT + Entity 方法增强了常规 BERT 的实体表示能力，使其能够学习更多的实体知识，需要大量数据支持，其有效训练严重依赖于大规模数据集。GENRE 方法通过自回归公式直接捕捉上下文与实体名称之间的关系，有效地实现了两者的交叉编码，但其依赖于预定义的候选实体集，并且在使用 Transformer 作为解码器时需要对 Wikipedia 摘要进行预训练。根据下表结果可以看出，本研究的模型相比于近年来的其他方法，F1 值提升了 1.8%~11.8%，验证了该模型的优越性和有效性。

Table 3. Comparison with existing methods on AIDA-CoNLL dataset
表 3. 与现有方法在 AIDA-CoNLL 数据集上的比较

Model	Micro-F1
BiLSTM + long range context attention [9]	82.4
KnowBert [15]	73.7
BERT + Entity [6]	79.3
GENRE [14]	83.7
AT-LOM	85.5

5. 结论

本研究通过对实体链接领域的研究现状进行综述并总结当前的实体链接模型主要存在三个缺点，提出了一种基于 Longformer 和更精细化模块设计的实体链接方法，并通过对多个主流方法的对比实验，验证了该方法在长文本实体链接任务中的优势。本研究结合 Longformer 强大的长文本处理能力和对比学习的优化机制，有效提升了模型在实体识别和链接方面的准确性与鲁棒性。通过精心设计的数据预处理和模型训练流程，系统地优化了模型的各项参数，达到了显著的性能提升。实验证明，在处理长文本时，该模型在实体链接任务中表现出色，相较于其他方法 Micro-F1 值取得了显著的提升。

本研究为实体链接领域的长文本处理提供了新的思路和有效的解决方案。然而，仍然存在一定的局限性，特别是在处理训练数据之外的领域或较冷门的实体时，模型的泛化能力仍有提升空间。未来的研究可以通过跨领域数据集的训练和微调，提升模型在不同场景下的适应性。同时，可以探索更加高效的预训练语言模型，并采用不同的训练模式，如少样本学习和迁移学习，以减少对大规模标注数据的依赖。此外，当前的方法主要基于文本信息进行实体链接，未充分考虑其他模态信息(如图像、视频等)在实体识别和消歧中的辅助作用。因此，未来可以探索多模态学习方法，结合图像、视频等信息，以进一步提升模型在特定任务中的表现。

基金项目

河北省社会科学基金(HB20TQ003)。

参考文献

[1] Chen, S., Wang, J., Jiang, F. and Lin, C. (2020) Improving Entity Linking by Modeling Latent Entity Type Information. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 7529-7537. https://doi.org/10.1609/aaai.v34i05.6251

[2] Sevgili, Ö., Shelmanov, A., Arkhipov, M., Panchenko, A. and Biemann, C. (2022) Neural Entity Linking: A Survey of Models Based on Deep Learning. *Semantic Web*, **13**, 527-570. https://doi.org/10.3233/sw-222986

[3] Shen, W., Wang, J. and Han, J. (2015) Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, **27**, 443-460. https://doi.org/10.1109/tkde.2014.2327028

[4] Cao, Y.X., Hou, L., Li, J.Z. and Liu, Z.Y. (2018) Neural Collective Entity Linking. *Proceedings of the* 27*th International Conference on Computational Linguistics*, Santa Fe, 20-26 August 2018, 675-686.

[5] Beltagy, I., Peters, M.E. and Cohan, A. (2020) Longformer: The Long-Document Transformer.

[6] Broscheit, S. (2019) Investigating Entity Knowledge in BERT with Simple Neural End-to-End Entity Linking. *Proceedings of the* 23*rd Conference on Computational Natural Language Learning* (*CoNLL*), Hong Kong, November 2019, 677-685. https://doi.org/10.18653/v1/k19-1063

[7] Young, T., Hazarika, D., Poria, S. and Cambria, E. (2018) Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. *IEEE Computational Intelligence Magazine*, **13**, 55-75. https://doi.org/10.1109/mci.2018.2840738

[8] De Cao, N., Aziz, W. and Titov, I. (2021) Highly Parallel Autoregressive Entity Linking with Discriminative Correction. *Proceedings of the* 2021 *Conference on Empirical Methods in Natural Language Processing*, November 2021, 7662-7669. https://doi.org/10.18653/v1/2021.emnlp-main.604

[9] Kolitsas, N., Ganea, O. and Hofmann, T. (2018) End-to-End Neural Entity Linking. *Proceedings of the* 22*nd Conference on Computational Natural Language Learning*, Brussels, 31 October-1 November 2018, 519-529. https://doi.org/10.18653/v1/k18-1050

[10] 李天然, 刘明童, 张玉洁, 等. 基于深度学习的实体链接研究综述[J]. 北京大学学报(自然科学版), 2021, 57(1): 91-98.

[11] Jia, B., Wang, C., Zhao, H. and Shi, L. (2022) An Entity Linking Algorithm Derived from Graph Convolutional Network and Contextualized Semantic Relevance. *Symmetry*, **14**, Article No. 2060. https://doi.org/10.3390/sym14102060

[12] Zhang, W., Hua, W. and Stratos, K. (2021) EntQA: Entity Linking as Question Answering.

[13] Poerner, N., Waltinger, U. and Schütze, H. (2020) E-BERT: Efficient-yet-Effective Entity Embeddings for BERT. *Findings of the Association for Computational Linguistics*: *EMNLP* 2020, November 2020, 803-818. https://doi.org/10.18653/v1/2020.findings-emnlp.71

[14] De Cao, N., Izacard, G., Riedel, S., *et al*. (2020) Autoregressive Entity Retrieval.

[15] Peters, M.E., Neumann, M., Logan, R., Schwartz, R., Joshi, V., Singh, S., *et al*. (2019) Knowledge Enhanced Contextual Word Representations. *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*), Hong Kong, November 2019, 43-54. https://doi.org/10.18653/v1/d19-1005

[16] Humeau, S., *et al*. (2019) Poly-Encoders: Architectures and Pre-Training Strategies for Fast and Accurate Multi-Sentence Scoring. *International Conference on Learning Representations*, New Orleans, 6-9 May 2019, pages.

[17] Tsai, C.T., Upadhyay, S. and Roth, D. (2025) Introduction to Entity Discovery and Linking. *Multilingual Entity Linking*, Springer Nature Switzerland, 1-14. https://doi.org/10.1007/978-3-031-74901-8_1

[18] Zhong, L., Wu, J., Li, Q., *et al.* (2023) A Comprehensive Survey on Automatic Knowledge Graph Construction. *ACM Computing Surveys*, **56**, 1-62. https://doi.org/10.1145/3618295

[19] Hoffart, J., Yosef, M.A., Bordino, I., *et al*. (2011) Robust Disambiguation of Named Entities in Text. *Proceedings of the* 2011 *Conference on Empirical Methods in Natural Language Processing*, Edinburgh, 27-29 July 2011, 782-792.

[20] Miyato, T., Dai, A.M. and Goodfellow, I. (2016) Adversarial Training Methods for Semi-Supervised Text Classification.