Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2024, 13(2), 425-440
Published Online May 2024 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2024.132044

Hans汉斯

探索中文文本中实体关系智能提取：一种基于数据与模型协同优化的新方法

刘翠媚，罗序良*，郭凤婵，吴毅良

广东电网有限责任公司江门供电局，广东 江门

收稿日期：2024年4月19日；录用日期：2024年5月24日；发布日期：2024年5月31日

摘　要

本文旨在解决从非结构化的中文文本中提取实体和关系的问题，重点关注命名实体识别(NER)和关系提取(RE)所面临的挑战。为了增强识别与提取能力，我们设计了一个管道模型，分别应用于NER和RE，并整合了外部词典信息以及中文语义信息。我们还引入了一种创新的NER模型，结合了中文拼音、字符和词语的特征。此外，我们利用实体距离、句子长度和词性等信息来提高关系提取的性能。本文经过深入研究数据、模型和推理算法之间的关联作用，以提高解决这一挑战的学习效率。通过与现有多个方法的实验结果对比，我们的模型取得了显著的成果。

关键词

命名实体识别，关系提取，深度学习，双向长短期记忆网络，注意力机制

Exploring Intelligent Entity Relationship Extraction in Chinese Text: A New Method Based on Data and Model Collaborative Optimization

Cuimei Liu, Xuliang Luo*, Fengchan Guo, Yiliang Wu

Guangdong Power Grid Co., Ltd. Jiangmen Power Supply Bureau, Jiangmen Guangdong

Received: Apr. 19th, 2024; accepted: May 24th, 2024; published: May 31st, 2024

*通讯作者。

文章引用: 刘翠媚, 罗序良, 郭凤婵, 吴毅良. 探索中文文本中实体关系智能提取: 一种基于数据与模型协同优化的新方法[J]. 人工智能与机器人研究, 2024, 13(2): 425-440. DOI: 10.12677/airr.2024.132044

Abstract

This paper aims to address the problem of extracting entities and relationships from unstructured Chinese text, focusing on the challenges faced in Named Entity Recognition (NER) and Relation Extraction (RE). To enhance recognition and extraction capabilities, we designed a pipeline model specifically for NER and RE, integrating external dictionary information as well as Chinese semantic information. We also introduced an innovative NER model that combines features of Chinese pinyin, characters, and words. Furthermore, we utilized information such as entity distance, sentence length, and part-of-speech to improve the performance of relation extraction. We delved into the interplay between data, models, and inference algorithms to improve the learning efficiency in tackling this challenge. Compared to existing methods, our model has achieved significant results.

Keywords

Named Entity Recognition, Relation Extraction, Deep Learning, BiLSTM, Attention Mechanism

1. 引言

信息提取是自然语言处理(NLP)任务中的重要组成部分，其目标是将非结构化的文本信息转化为有价值且有组织的结构化信息。在信息提取任务中，命名实体识别(NER)、实体关系提取(RE)和事件提取是关键任务。命名实体识别的目标是在文本中识别实体元素并将其分类到预定义的实体类别中[1]。目前，命名实体识别可以通过基于规则的方法、基于统计机器学习的方法和基于深度学习的方法来实现[2]。实体关系提取常使用基于深度学习的模型，这是因为深度学习具有记忆训练模型和发现深层特征的能力[3]。迁移学习可以将一个已完成训练的网络权重迁移到其他新的网络，使其学习到原网络的特征。目前，中文命名实体识别的研究主要集中在基于字符识别的方法上。然而，在处理复杂的实体关系时，中文实体关系提取任务仍然面临着巨大的挑战。主要的挑战有以下三点：

1)、如何引入实体字符之外的词汇：中文句子中的语义是由词语构成的，这意味着中文词语包含丰富的信息[4]。通过学习大量未标注文本语料库中的词语嵌入，可以降低在文本语料库中遇到未知词语的影响。引入外部词汇还有助于更好地解决词语边界不清的问题。

2)、如何更好地利用中文中的丰富语义信息：深度学习领域广泛研究了基于字符、词语属性、发音等特征的模型，以提高关系提取的准确性和泛化能力。充分利用语义信息有助于神经网络更好地提取实体之间的关系。

3)、数据分布的对称性一致性：在学术讨论中，确保数据、模型和推理算法三元组内的对称性与数据分布的对称性保持一致对于实现最佳学习效率至关重要。解决这个问题需要深入探讨这些对称性之间的关系，并开发更灵活的模型以提高学习效率。

为了解决以上问题，本文使用流水线模型分别对命名实体识别模型和关系提取模型进行建模，并在相关数据集上取得了良好的结果。本文的主要贡献包括：引入了一种新颖的 NER 模型，结合中文拼音、

字符和词语以增强识别能力。融合实体距离、句子长度和词性信息提高了关系提取模型的性能。探索了数据、模型和推理算法之间的相互作用，研究了深度学习中的协同效应和对称性，从而提高了学习效率。

在引言的最后一段，我们提供了论文结构的概述以便于理解。具体来说：第 2 节深入研究相关工作，包括信息提取研究现状、预训练模型、解码分类层和(D, M, I)三元组的研究工作。第 3 节详细阐述了本文所提出的模型，包括基于字符、词语和拼音的命名实体识别模型，以及基于多特征绑定的关系提取模型。第 4 节对实验结果进行分析，分别对本文提出的命名识别模型与关系提取模型进行了性能评估。第 5 节提供了对实验结果的深入分析，重点分析了命名实体识别(NER)和关系提取(RE)模型的性能表现。最后，第 6 节总结了本文的研究成果并讨论了未来的研究方向。

2. 相关工作

2.1. 信息提取的研究现状

统计机器学习算法的出现开启了命名实体识别算法的新时代，该任务现已与神经网络相结合。研究人员开始使用基于字符嵌入和词语嵌入的循环神经网络(RNN)识别句子中的命名实体，解决了传统统计方法中存在的特征工程问题。随着深度学习的不断发展，利用深度学习方法解决命名实体识别问题已成为当前热门的研究课题。这类方法的优势在于神经网络模型可以自动学习句子特征，无需繁琐的特征计算工程[5]。

在基于词向量的处理中，文本首先被分割成词语，然后将词语映射成向量形式，即将每个词语表示为一个特征向量。这种方法面临的主要问题包括词语分割错误导致命名实体被错误切割、词汇表维度过大以及无法有效表征相似词语之间的关联[6]。其优点在于词语边界信息和语义特征清晰。

命名实体识别中的双向长短期记忆网络(BiLSTM)通过考虑每个词语的上下文，实现了全面的序列建模，从而提高了命名实体识别的准确性。该模型能够提取序列特征、处理长距离依赖关系并生成动态序列表示，为模型提供了灵活性。BiLSTM 通过学习高级语言特征，在 NER 任务中表现出色，已成为有效的序列标注方法。

后来，基于特征提取的方法被用于通过输入实体及其相应的文本、句子或语言相关特征对关系进行分类的任务上。上述方法的有效性在很大程度上取决于现有自然语言处理工具中派生的特征质量[7]。神经网络的出现为特征提取开辟了新方向。卷积神经网络(CNN)、循环神经网络、图神经网络和长短期记忆网络等模型被广泛应用于命名实体识别和关系提取任务中。在 NLP 任务中，文本输入通常转化为信息序列，由于 RNN 适合处理序列信息，因此在 NLP 任务中得到广泛应用。图神经网络适用于利用点与点集之间的关系[8]，能够有效捕捉文本的结构信息，近年来也受到自然语言处理任务的关注。

Daojian Zeng 等人[9]提出的分段卷积神经网络(PCNN)方法在训练数据相对较少的情况下采用多样本学习方法，允许标记句子或示例中存在误标记。此外，使用分段最大池化卷积架构自动学习相关特征也取得了较好的效果。

2.2. 预训练模型

预训练语言模型通常基于 Transformer 架构，例如 BERT [10]。因 BERT 在中文序列标注任务中性能优异，近年来众多研究人员采用了不同的策略来利用 BERT 的优势。Yang 等人[11]在 BERT 的基础上添加了一个简单的 softmax 层，在中文词语分割(CWS)任务中取得了最优异的性能表现。这表明 BERT 的上下文表示对中文词语分割任务至关重要，简单的分类层就足以带来显著的性能提升。此外，Meng 等人[12]的研究表明，在 BERT 中利用字符特征可以显著超越基于静态嵌入的方法，特别是在中文命名实体识别(NER)和中文词性(POS)标注任务中。这表明 BERT 不仅提高了整体性能，而且其字符级信息对于中文文

本的结构化任务也很有价值。利用外部词典特征来增强字符特征已被证明是有效的，如 SoftLexicon [13]和 LEBERT [14]所示。在 SoftLexicon 中，词典特征被合并到字符表示中，避免了使用复杂架构整合词典特征的必要性。LEBERT 基于此方法，将词典特征集成到 BERT 的较低层中。BERT 的较低层促进了词典特征和 BERT 之间的更深层次交互。LEBERT 的主要思想是将来自 BERT 的上下文表示和词典特征集成到中文 NER 模型中。

本文修改了 LEBERT 的嵌入模型，用轻量级且更强大的 Mengzi 模型[15]取代了它。与相同大小甚至更大维度的模型相比，Mengzi 模型表现出显著的性能提升。修改预训练模型和微调策略已被证明可以有效提升基线结果。在不修改底层模型架构的情况下，Mengzi 模型是目前最先进的中文预训练语言模型之一。

2.3. 解码分类层

当前主流的分类器主要有条件随机场(CRF)解码分类器[16]和 softmax 解码分类器[17]。CRF 是一种无向图模型，用于对由随机变量组成的序列进行联合概率分布建模，常用于序列标注、自然语言处理和计算机视觉等领域。CRF 可以通过学习标签之间的依赖关系和句子中的约束关系来获得标签的全局最优序列。在序列标注任务中，需要对给定输入序列的每个位置进行标注，这可转化为对每个位置的标注建模为一系列随机变量。CRF 通过在这些变量之间构建概率图模型来表示它们的相互依赖关系，即给定一个位置的标注，其他位置标注的条件概率将被建模。CRF 模型通过考虑相邻位置的标注来构建全局组合概率分布。CRF 模型的训练通常采用最大似然估计方法，通过最大化训练数据的似然函数来得到每个特征函数的权重。综上所述，CRF 模型通过建立全局联合概率分布描述标注序列的生成过程，并通过特征函数对标注和序列之间的关系进行建模。

另一方面，softmax 的作用是将输入序列映射到命名实体分类标签的概率分布并通过贪心算法让每个状态获得概率最大的分类标签，适合用于多分类问题的处理。具体来说，softmax 函数充当分类器，为每个输出的分类结匹配一个概率值，反映其属于对应类别的可能性，而不仅仅是确定一个最大值。在本文中，BiLSTM 之后使用 CRF 做解码层，以提升模型预测的准确性。

2.4. (D, M, I) 三元组

Lechao Xiao 等人[18]提出了一种综合系统，将机器学习视为数据、模型和推理算法的结合。从对称性的角度研究了机器学习中的三元组(数据、模型、推理算法)不同组合与性能之间的关系。当算法与数据的固有分布相一致时，机器学习的学习效率最有优势。这种对称性使算法能够识别数据中的相关模式和关系，从而提高机器学习的性能和泛化能力。(D, M, I)三元组的核心内容表述如下：

1). 数据(D)：

数据是神经网络模型的基础，模型的有效性在很大程度上取决于所使用数据的质量和数量。数据增强在自然语言处理任务中尤为重要，因为自然语言数据的复杂性和多样性使得模型难以完全理解语言的真实含义。选择高质量的词嵌入对于构建高性能自然语言处理模型是非常重要的。高质量的词嵌入可以提高模型对词语的准确把握，并增加模型对新数据源的适应性。此外，词嵌入的数量对模型的性能有显著影响，更多的词嵌入通常允许模型学习更多特征。然而，仅依赖更多的词嵌入是不够的，词嵌入的质量也很重要。如果嵌入包含错误、噪声或不一致性，这些问题将被模型学习，并会影响模型的性能。因此，为了构建高性能的自然语言处理模型，有必要选择合适规模的词嵌入数据集，并仔细清理和处理数据，以确保数据的质量和数量都能最大化模型的性能。

2). 模型(M)：

设计最优化数据效率的神经网络模型对于解决自然语言处理任务至关重要。在建模框架中，模型选

择很重要，确定适合任务的模型需要考虑多个方面。在本文中，我们旨在分析适用于命名实体识别和关系提取任务的长短期记忆网络(LSTM)模型和门控循环单元(GRU)模型的优缺点。与LSTM模型相比，GRU模型具有更少的参数和更快的训练速度，但在处理更长的序列时，LSTM 表现更好。因此，在选择合适的模型时，需要考虑数据和模型之间的关联性和协同作用。

3). 推理算法(I)：

推理算法指的是神经网络模型执行学习过程的方法。强大的机器学习性能可能来自(M, I)、(D, I)或(D, M, I)之间不同组合的相互作用。不同的推理算法需要针对不同的任务进行选择不同的组合模型。在我们的研究中，我们通过实验表明，不同的推理算法在不同模型的情况下有不同的性能。

3. 模型设计

3.1. 基于字符、词语和拼音组合的命名实体识别模型

本研究采用字符级输入作为模型输入，这导致某些特征受到模糊边界的影响。为了解决这一问题，本文引入词嵌入以增强命名实体识别的准确性。通过利用词嵌入，模型能够更好地捕捉词语之间的语义关联，并根据上下文理解词语的含义。中文命名实体识别模型建立在 LEBERT 模型之上，该模型整合了超过 1200 万个中文词语的语义信息，这些词语在大规模、高质量数据集上进行了预训练[19]。此外，该模型还利用了中文字符独特的拼音向量，使用双向长短期记忆网络(BiLSTM)对字符序列进行标注预测，并最终采用条件随机场(CRF)优化预测标注。与原始模型相比，该模型在多个中文命名实体识别数据集上的表现均优于原始模型，显示出对不同类别的识别率显著提高，并且每个标注的准确率也有所提升。模型架构如图 1 所示。

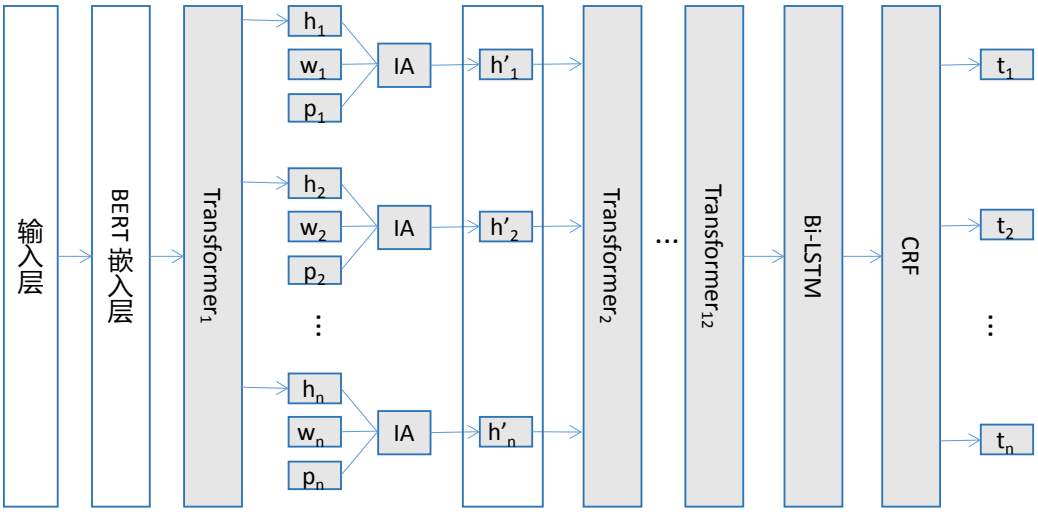

Figure 1. Named entity recognition model architecture based on the combination of Chinese characters, words, and pinyin

图 1. 基于汉字、词、拼音组合的命名实体识别模型架构

在利用拼音嵌入时，本文采用基于 ChineseBERT 的方法生成拼音嵌入。首先，使用开源的 pypinyin 包为每个字符获取其拼音序列。然后，对每个字符的拼音进行填充，使其长度一致。使用卷积神经网络(CNN)模型处理该序列，并通过最大池化得到最终的序列嵌入。这种方式保证了输出向量的维度不受输入序列长度的影响。输入序列的长度固定为 8。当序列长度小于 8 时，剩余的位置使用特殊字符“-”进行填充，如图 2 所示。

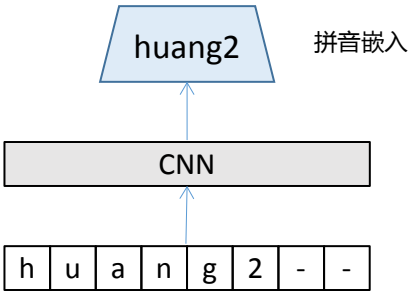

Figure 2. Generate pinyin embeddings
图 2. 拼音嵌入的生成

针对给定的输入文本序列，经过预处理步骤后，序列被输入到 BERT 的嵌入层，该层为文本序列生成了标记嵌入、位置嵌入和段落嵌入。随后，句子被输入到 BERT 的初始 Transformer 层，产生词语部分特征 $h_1, h_2, \cdots, h_n$。接着，通过预训练词嵌入查找表(例如腾讯 AI 实验室嵌入语料库)获取词语集嵌入向量 $w_i = \{w_{i1}, w_{i2}, \cdots, w_{in}\}$，其中 w_i 表示第 i 个字符的词语嵌入向量集。词语嵌入 w_{i1}、w_{i2} 等的维度可能与模型中字符向量 h_i 的维度不同。为了确保字符向量 w_{ij} 与字符向量 h_i 之间的对齐，需要对词语向量进行非线性变换，如公式(1)所示：

$$x_{ij} = W_2\left(\tanh\left(W_1 w_{ij} + b_1\right)\right) + b_2 \tag{1}$$

其中，W_1 是一个 $d_c \times d_w$ 的矩阵，W_2 是一个 $d_c \times d_c$ 的矩阵，b_1 和 b_2 是标量偏差。d_c 表示 BERT 的隐藏大小，d_w 表示词语嵌入的维度。上述参数与 tanh 函数一起实现了词语嵌入的转换。受 LEBERT 的启发，为了将词典特征和拼音特征与字符信息相结合，我们设计了一个新颖的信息适配器(IA)，如图 3 所示。

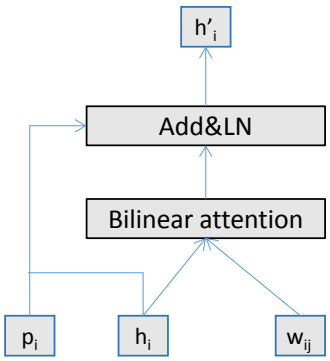

Figure 3. Information adapters
图 3. 信息适配器

信息适配器采用注意力机制将字符特征和词典特征集成到 BERT 中，并进一步将拼音信息纳入适配器。我们将拼音嵌入向量 p_i、字符嵌入向量 h_i 和词语的词语嵌入向量 x_i 输入到信息适配器层中。为了计算每个匹配词语的相关权重，我们借鉴了 LEBERT 的字符-词语注意力机制。具体来说，我们将第 i 个字符向量表示为 h_i，并将与第 i 个字符相对应的词语嵌入向量集表示为 x_i，其中 x_{ij} 表示与第 i 个字符相对应的第 j 个词语的嵌入。我们将所有 x_i 映射到 h_i 上，即 $x_i = \{x_{i1}, x_{i2}, \cdots, x_{im}\}$，维度为 m 乘以 d_c。每个词语的相关性计算如公式(2)：

$$A_i = \text{softmax}\left(h_i W_{attn} x_{ij}^T\right) \tag{2}$$

其中W_{attn}是双线性注意力(Bilinear Attention)的权重矩阵，$A_i=\{\alpha_{i1},\alpha_{i2},\cdots,\alpha_{in}\}$是每个词语的权重向量。这样，我们可以得到公式(3)所示的词语特征的加权和。

$$h_i'=\sum_{j=1}^{m}\alpha_{ij}x_{ij} \tag{3}$$

通过公式(4)的方法将加权的词语嵌入和拼音特征注入到字符向量中：

$$h_i'=h_i+h_i'+p_i \tag{4}$$

将加权的词语嵌入和拼音特征注入到字符向量中，采用 dropout 层和层归一化进行整合。所获得的信息经过一系列 11 个 Transformer 块的处理，得到 BERT 的最终输出。该输出随后被输入到双向 LSTM 中，从而获取输入向量的序列特征值。最后，通过条件随机场(CRF)得到最终的预测结果输出。我们将改造的命名实体识别模型命名为 LBC (LEBERT + BiLSTM + CRF)模型。

3.2. 基于多特征绑定的关系提取模型

为了提升中文文本关系提取的准确性，本文基于 BERT 编码器获取更多语义特征信息，充分利用包含在 BERT 模型中的信息。本文采用基于 BERT 模型的中文预训练模型对文本关系进行预测。通过对数据集中的数据进行分析后发现，大多数句子的关系包含在两个实体之间。如果两个实体彼此靠近，且之间没有其他关系，则目标关系可能隐藏在第二个实体之后。例如，在新闻文章中提及两个人之间的关系时，这些人的姓名可能紧密相关。通过添加两个实体之间的距离信息(以单词或字符的数量表示)作为特征，可以更好地确定实体关系信息的位置。

此外，两个实体之间的关系通常与句子的长度有关。因此，句子长度信息可以帮助识别不同长度文本之间的差异。例如，在较短的文本中，关系可能更容易识别，因为干扰信息较少；而在较长的文本中，关系可能更难识别，因为干扰信息更多。因此，长度特征可以提供有用的线索，帮助模型更准确地执行关系提取。中文的词性属性通常包含有丰富的特征信息。本文使用 THULAC [20]工具包将句子分词并提取词性属性。然后将这些属性与其他词性属性(例如形容词和介词)进行比较。一般来说，名词或动词相比其他类别的词性更有可能包含关系信息。例如，在句子“刘备、关羽和张飞宣誓兄弟情谊，成为兄弟”中，THULAC 的词性标注结果为：“刘备/名词，关羽/名词，和/连词，张飞/名词，宣誓兄弟情谊/动词，成为/动词，兄弟/名词”。其中，“宣誓兄弟情谊”为动词，“兄弟”为名词，表明每对实体之间存在关系。

最后，将多个特征整合在一起并输入模型。实验结果表明，该方法在多个中文关系提取数据集上的 F1 值有所提高，不同类别关系的识别率也得到了显著提升。此外，模型通过引入预训练的外部词向量以获得了更多特征信息。其中第i个词语的词向量为V_i^w，通过使用预训练的词向量查找表获得；其他特征向量为V_i^{wj}，表示第 j 个特征。最终的词语特征向量表示如公式(5)。

$$X_i=\left[V_i^w,V_i^{w1},\cdots,V_i^{wm}\right] \tag{5}$$

其中，m表示除了词向量之外，还有m种其他特征。本文中，$m=3$。对于句子级文本序列，直接将词语标注信息输入双向 LSTM 进行编码，使用 F (前向)和 B (后向)表示两个方向。h_i (隐藏)和c_i (记忆)分别表示隐藏信息和全局信息；则在第i时刻的输出为：

$$F_i=\left[F_{hi},F_{ci},B_{hi},B_{ci}\right] \tag{6}$$

将句子级特征信息和词语级特征信息提取并拼接起来，形成最终的提取特征向量。词语级特征信息主要有 2 个实体：实体 1 (N1)和实体 2 (N2)。本文中，从特征嵌入和双向 GRU 获得的向量被拼接起来，

表示两个实体为$[X_{n1}, F_{n1}, X_{n2}, F_{n2}]$。句子级特征信息关注上下文信息，这是从双向GRU层的输出构建的，如图4所示。

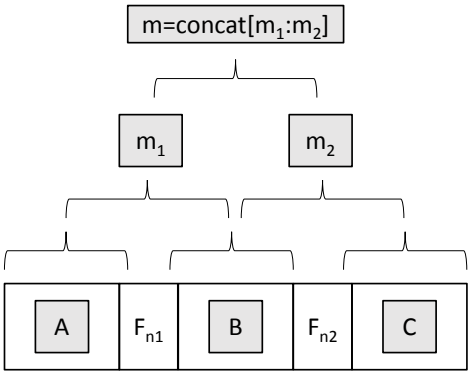

Figure 4. Construction of sentence-level feature vectors
图4. 句子级特征向量的构建

模型框架如图5所示。BiGRU输出的矩阵可划分为A、B和C三部分，由n1和n2构成。通过最大池化操作提取向量m1和m2，并将其连接形成输出信息。该输出随后输入注意力层进行加权求和。注意力机制从大量信息中过滤掉冗余信息，使模型可以更关注所需信息，从而增强其关注更重要特征的能力[21]。

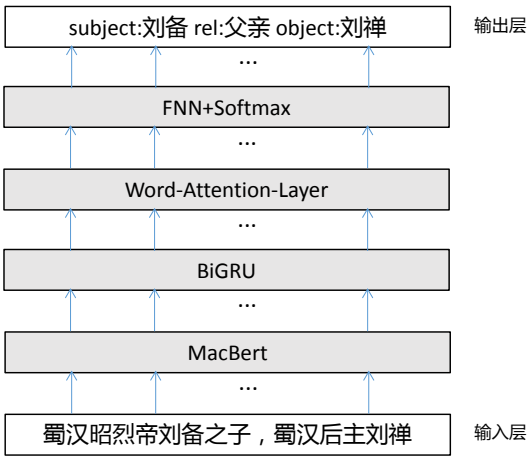

Figure 5. Framework diagram of the relational extraction model
图5. 关系提取模型框架图

本文基于训练于中文数据集上的BERT模型进行改造。BERT利用大量未标注数据进行预训练，使其能够更好地理解文本中的语境。这有助于识别关系提取任务中实体之间的具体关系，例如主谓关系或宾语关系。在预训练过程中，该模型学习了丰富的文本表示，并能够将实体和关系表示为高质量的向量表示。这些向量表示可用于各种关系提取任务，如实体分类和关系提取。由于预训练模型是在大规模数据上训练的，因此它学习了丰富的文本表示和语言知识。这使得该模型能够在没有大量标注数据的情况下学习新类别的关系。本文将改进的关系提取模型命名为BBA (BERT + BiGRU + Attention)模型。

图5中的输入层包含一个中文句子，其意思为：“蜀汉后主刘禅是刘备之子”。在输出层中，三元

组的含义如下：对象：刘备，关系：父子关系，对象：刘禅。

4. 实验结果及分析

为了验证本文所提出的命名实体识别模型 LBC 与关系提取模型 BBA 的性能表现，分别对两个模型进行实验数据分析。在本研究中，实验环境基于 PyTorch 框架构建，具体的实验环境配置如表 1 所示。

Table 1. Configure the experimental environment
表 1. 实验环境配置

配置名	版本
CPU	Intel Core i9-13900K 24C32T
GPU	NVIDIA GeForce RTX 4090 24G
RAM	128 GB
SSD	1 TB
Python	3.9
Pytorch	1.12.1
CUDA	11.6
操作系统	Ubuntu 20.04.6 LTS

4.1. 命名实体识别模型 LBC

为了验证 LBC 模型在命名实体识别任务中的有效性，本文使用了三个常用的中文命名实体识别数据集。每个数据集都按比例被划分为训练集、验证集和测试集。模型的性能使用 F1 分数作为评估指标进行评估，并与其他主流的中文命名实体识别模型进行比较。

4.1.1. 数据集

本文使用三个常用的中文命名实体识别数据集：微博[22]、简历[23]和 MSRA [24]。微博数据集的语料库来自社交媒体，包含四种实体类型：个人、地点、组织和地缘政治实体。MSRA-NER 数据集由微软亚洲研究院发布，包含 3 类实体：人名、地名和组织名。简历数据集是中国股市上市公司高管简历的集合，包含 3 类实体：人名、地名和职位。

4.1.2. 参数配置

本文中的命名实体识别模型参数经多次实验后进行了优化调整，最终模型参数配置为表 2 所示。

Table 2. Named entity recognition model parameters
表 2. 命名实体识别模型参数

参数项	参数值
Max Length	150
Crf_lr	1×10^{-4}
Adapter_lr	1×10^{-4}
Epoch	20
Batch Size	12
Loss_Type	Ce
Learning Rate	1×10^{-4}
dropout	0.5

4.1.3. 实验结果

通过表 3 显示的实验结果可知，本文改进的实体识别模型 LEBERT-BiLSTM 在简历、微博和 MSRA 三个数据集的综合性能上显著优于其他 5 个模型。另外，从对比结果中可看出数据集越小，模型效果提升越明显。

Table 3. Comparison of F1 scores of multiple Chinese named entity recognition datasets
表 3. 多个中文命名实体识别数据集的 F1 分数对比

参数项	F1 (%)		
	简历	微博	MSRA
BERT	95.32	67.22	94.74
ERNIE	94.81	67.94	95.03
Chinese BERT	/	70.81	/
MFE-NER	95.75	67.71	89.94
LEBERT	96.03	70.74	95.71
LEBERT-BiLSTM	96.64	73.78	95.87

如表 3 所示，三个数据集的 F1 分数均有所提升，表明融合词向量和拼音向量以提高命名实体识别的准确性是有效的。表 3 前五行为基于 BERT 模型的变体模型，第一行为 BERT 基线模型。第二行为百度于 2019 年 4 月基于 BERT 模型优化的 ERNIE 模型[25]。其后是 Chinese BERT [26]，其整合了中文字符的字形信息和拼音信息，并融合了中文字符、字形、发音和上下文之间的联系。字形向量由不同的部件构成，拼音向量则由相应的罗马化拼音字符序列构成。二者与词向量融合后，作为预训练模型的输入序列。第四行为 MFE-NER，其采用了类似的融合字形和拼音的方法，但采用了独特的字体处理方式。该方法捕捉字形特征，并使用“五笔”编码方案表示中文字符的结构模式，增强了嵌入空间中具有相似字形结构的字符的邻近性。最后，原始的 LEBERT 模型将词典信息输入模型，以增强命名实体识别能力。本文在 LEBERT 模型基础上进行了改造，并与多个模型进行了实验比较。结果表明，本文改进的中文语言模型与词典特征相结合进行中文命名实体识别是有效的。与 LEBERT 模型相比，本文模型在不同类别实体的识别上均有大幅提升。微博数据集上不同类别实体识别的 F1 值对比如图 6 所示。

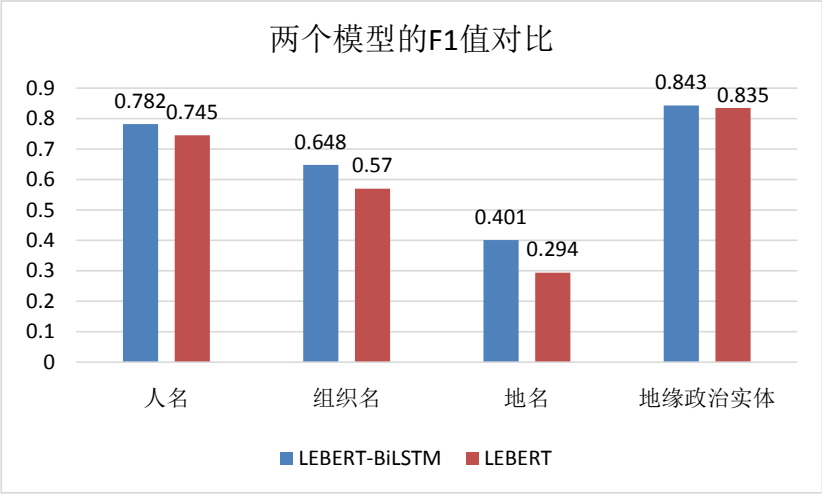

Figure 6. F1 values of named entities with different labels in the Weibo dataset
图 6. 微博数据集下不同标签的命名实体 F1 值

如图 6 所示，本文改进的 LEBERT-BiLSTM 模型相对于 LEBERT 模型在几个实体分类中的识别效果显著提高，特别是对于特定地名和特定组织名。

4.1.4. 消融实验

为进一步证明本文模型 LBC 的优越性能，本文基于 LEBERT 修改模型进行了多个消融实验，实验结果如表 4 所示。在 LBC 模型中，BiLSTM 对预训练模型编码的输入序列的向量进行序列建模，有效地提取连续的序列信息。这种连续的序列信息对于命名实体识别非常有效，并在相应数据集上取得了良好的结果。

Table 4. Experimental results of different models under the Weibo dataset
表 4. 微博数据集下不同模型的实验结果

模型	F1 (%)
LEBERT-CRF	70.73
LEBERT-LSTM-CRF	71.87
LEBERT-BiGRU-CRF	71.13
LEBERT-BiLSTM-Softmax	72.62
LBC (LEBERT-BiLSTM-CRF)	73.75

4.2. 关系提取模型 BBA

本研究采用了一种基于 BERT 和双向门控循环单元(Bi-GRU)的模型进行中文关系提取。首先，使用 BERT 模型对输入文本进行编码，生成一个向量序列。然后，将该序列输入到 Bi-GRU 模型中，得到 Bi-GRU 模型的输出。接着，通过注意力层对 Bi-GRU 模型的输出进行加权，得到一个加权和作为最终输出向量。最后，将输出向量输入到线性层，并通过映射生成与每个关系相关的概率。

为了验证模型的有效性，本本文分别在三个中文关系提取数据集上进行了实验。数据集按照原始确定的比例划分为训练集、测试集和验证集。实验采用准确率、召回率和 F1 值作为评价指标，对模型的中文关系提取性能进行了评估，并与其他中文关系提取模型进行了比较。

4.2.1. 数据集

本文使用了两个中文关系提取数据集：DuIE 中文关系提取数据集[27]和通过修改获得的中文文学文本话语级关系提取数据集(Chinese-Literature-RE-Dataset) [28]。表 5 展示了这两个数据集的划分比例。

对于 DuIE 数据集，我们首先从数据集中提取出实体 A、关系、实体 B 以及包含实体的文本。然后，我们对这些数据进行格式化和标准化，转换为 JSON 格式。我们手动删除了 2164 个句子，这些句子在清理数据时由于文本中存在多个双引号而导致 JSON 文件格式错误。最终，我们得到了 362,516 个 DuIE 训练数据集、50,000 个测试集数据和 45,429 个验证集数据。

Table 5. The number of DuIE datasets and Chinese-Literature-RE-Dataset datasets
表 5. DuIE 数据集和 Chinese-Literature-RE-Dataset 数据集的数量

数据集	DuIE	Chinese-Literature-RE-Dataset
训练集数量	362,516	19,447
测试集数量	50,000	2220
验证集数量	45,429	2220
关系类型数量	49	10

对于 Chinese-Literature-RE-Dataset 数据集，由于该数据集的文本是一篇文学作品，我们首先通过句号将文章文本分割成句子。然后，我们从相应的文本文件中提取实体和关系注释，并将其标准化为新的 JSON 文件。总共我们获得了 19,447 个训练数据集、2220 个测试集和 2220 个验证集数据。

4.2.2. 实验参数设置

同一模型在不同的超参数设置下性能表现有所不同。因此，进行多次实验是必要的，使用多种不同的参数设置，并根据实验结果选择性能更好的模型参数作为最终的参数。本文中的最优参数设置如表 6 所示。

Table 6. Experimental parameters of the relational extraction model
表 6. 关系提取模型的实验参数

参数项	参数值
Torch.Size	[x, 1, 128]
Self Attention Hidden layer dimension	768
Dropout	0.5
Learning Rate	0.001
Epoch	10
Batch Size	8

4.2.3. 实验结果

为了验证本文提出的 BBA 关系提取模型的有效性，本文分别在 DuIE 数据集与 Chinese-Literature-RE-Dataset 数据集上进行了多个模型的对比实验。各模型的实验比较结果如表 7 和表 8 所示。

从表 7 和表 8 中的实验结果可以看出，本文提出的 BBA 模型在 DuIE 和 Chinese-Literature-RE-Dataset 关系提取数据集上均提高了准确率、召回率和 F1 值。另外，结果还显示对于关系数量较少的数据集，提升效果更为显著。

Table 7. Experimental results on the DuIE dataset
表 7. DuIE 数据集下的实验结果

模型	评估数据		
	准确率(%)	召回率(%)	F1 分数(%)
BiLSTM-Attention	86.16	85.27	85.61
PCNN-Attention	87.38	87.13	87.23
BERT	94.21	92.72	93.42
BBA	94.62	94.65	94.64

Table 8. Experimental results on the Chinese-Literature-RE-Dataset dataset
表 8. Chinese-Literature-RE-Dataset 数据集下的实验结果

模型	评估数据		
	准确率(%)	召回率(%)	F1 分数(%)
BiLSTM-Attention	88.91	88.22	88.56
PCNN-Attention	95.72	67.75	89.95
BERT	91.73	90.64	91.21
BBA	93.36	91.79	92.53

实验结果表明，BBA 模型在中文关系提取任务中表现良好。然而，在中文关系提取任务中的良好性能并不意味着该模型在其他数据集上也能表现同样出色。因此，下一步的研究工作需要对该模型进行全面的评估，以确定其在其他数据集上的实际效果。此外，需要持续优化模型的性能，以满足中文关系提取任务不断变化的需求。

5. 讨论分析

本文所提出的实体识别模型 LBC 和关系提取模型 BBA 结构简单，实验结果表明，本文改进的模型在相关数据集上取得了较好的效果。本节将重点探讨影响模型性能的相关因素。

5.1. 命名实体识别模型

通过融合字向量、拼音向量和词向量，显著提升了命名实体识别模型的性能。表 9 展示了相同配置环境下不同向量融合的实验结果。

Table 9. Experimental results after the fusion of different vectors
表 9. 不同向量融合后的实验结果

向量组合	F1 (%)
拼音 + 字	69.51
拼音 + 字 + 词向量	70.83
拼音 + 词向量	68.31
拼音 + 词向量(1200 万词向量)	72.14
拼音 + 词向量(200 万词向量)	73.75

本研究利用中文字形特征，将中文文本分解为字、词和拼音。实验表明，整合更多的特征不一定能带来更好的结果。Jordan Hoffmann 等人[29]指出，模型大小应与训练数据规模相匹配才能得到更好的效果。实验结果表明，相比于小型的 200 万词向量，1200 万词的大词向量不一定能带来更好的效果。当训练数据集较小时，小词向量可能比大词向量更有效。这是因为大词向量通常包含大量的冗余信息，这些信息对于解决小规模任务并不必要，反而可能干扰模型学习。这进一步强调了三元组系统中数据平衡的重要性，因为精心策划的数据集有利于提高自然语言处理任务的效率。

在自然语言处理任务中，向量维度通常包含语料库中字符的特征数量。高维向量通常能提取到更丰富的信息，对自然语言处理任务更有益。如表 10 所示，使用 200 维向量有助于提升命名实体识别的性能。

Table 10. Experimental results of word vector embedding in different dimensions on the Weibo dataset
表 10. 微博数据集上不同维度词向量嵌入的实验结果

向量组合	F1 (%)
拼音 + 词向量(1200 万词向量) (100 维)	67.43
拼音 + 词向量(200 万词向量) (100 维)	68.11
拼音 + 词向量(1200 万词向量) (200 维)	72.12
拼音 + 词向量(200 万词向量) (200 维)	73.74

5.2. 关系提取模型

关于关系提取任务，本文测试了词嵌入向量的维度，结果如表 11 所示。结果表明，对于自然语言处

理任务，词向量的维度越高，处理效果越好。选择与模型和推理算法数据类型相匹配的词嵌入向量，有助于提升模型性能。

Table 11. Experimental results of different dimensional word embeddings in the Chinese-Literature-RE-Dataset
表 11. Chinese-Literature-RE-Dataset 数据集中不同维度词嵌入的实验结果

模型	F1 (%)
BBA (200 维)	87.53
BBA (100 维)	92.52

综上所述，各种尝试已证明本文研究有助于提升自然语言处理任务的性能。本文着重探讨了词向量的维度，而没有深入研究替代不同维度其他层的可能性是否能带来更好的效果。在未来的工作中，我们将积极探索更多替代方案，以进一步提高模型的性能。

6. 结论

本文提出了一种新的中文文本中实体关系智能提取方法，通过数据与模型的协同优化来解决中文命名实体识别(NER)和关系提取(RE)中存在的挑战。研究中引入了一种新颖的 NER 模型，该模型结合了中文拼音、字符和词语，以增强对实体的识别能力。同时，关系提取模型通过融合实体距离、句子长度和词性信息，提高了性能。此外，文章还探讨了数据、模型和推理算法之间的相互作用，并研究了深度学习中的协同效应和对称性，这些研究有助于提高学习效率。实验结果表明，所提出的 LBC 模型和 BBA 模型在多个中文 NER 和 RE 数据集上均取得了良好的性能，优于现有的一些主流模型。消融实验的结果进一步证明了模型中各个组件的有效性。此外，论文还讨论了影响模型性能的因素，如词向量的维度和数据集的规模。

本研究在中文实体关系提取领域取得了显著进展，但仍面临诸多挑战，未来的研究需着重提升模型的泛化能力，使其能更好地适应跨领域和跨语言的应用场景；同时，优化计算效率，减少计算资源消耗，加快训练与推理速度；改进特征融合策略，更有效地整合多维度语言信息；探索小样本学习模型，解决数据标注不足的问题；采用多任务学习框架，实现任务间的数据共享与性能提升；增强模型的可解释性，开发可视化工具以揭示决策过程；并提高模型对噪声和异常数据的鲁棒性，以适应真实世界的数据不完美性。通过这些努力，预期将推动中文实体关系提取技术向更高水平发展。

基金项目

本文由“南网高层次人才特殊支持计划”项目资助。

参考文献

[1] 焦凯楠, 李欣, 朱容辰. 中文领域命名实体识别综述[J]. 计算机工程与应用, 2021, 57(16): 1-15.

[2] 刘浏, 王东波. 命名实体识别研究综述[J]. 情报学报, 2018, 37(3): 329-340.

[3] 康怡琳, 孙璐冰, 朱容波, 等. 深度学习中文命名实体识别研究综述[J]. 华中科技大学学报(自然科学版), 2022, 50(11): 44-53. https://doi.org/10.13245/j.hust.221104

[4] 钟诗胜, 陈曦, 赵明航, 等. 引入词集级注意力机制的中文命名实体识别方法[J]. 吉林大学学报(工学版), 2022, 52(5): 1098-1105. https://doi.org/10.13229/j.cnki.jdxbgxb20200984

[5] 何玉洁, 杜方, 史英杰, 等. 基于深度学习的命名实体识别研究综述[J]. 计算机工程与应用, 2021, 57(11): 21-36.

[6] 谢文芮. 融合字符多语义特征的命名实体识别研究与实现[D]: [硕士学位论文]. 无锡: 江南大学, 2022. https://doi.org/10.27169/d.cnki.gwqgu.2022.000329

[7] Cui, M., Li, L., Wang, Z., *et al*. (2017) A Survey on Relation Extraction. *Language, Knowledge, and Intelligence*: *Second China Conference*, *CCKS* 2017, Chengdu, 26-29 August 2017, 50-58.

[8] Xiong, S., Li, B. and Zhu, S. (2023) DCGNN: A Single-Stage 3D Object Detection Network Based on Density Clustering and Graph Neural Network. *Complex & Intelligent Systems*, **9**, 3399-3408. https://doi.org/10.1007/s40747-022-00926-z

[9] Zeng, D., Liu, K., Chen, Y. and Zhao, J. (2015) Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. *Proceedings of the* 2015 *Conference on Empirical Methods in Natural Language Processing*, Lisbon, 17-21 September 2015, 1753-1762. https://doi.org/10.18653/v1/D15-1203

[10] Devlin, J., Chang, M.W., Lee, K., *et al*. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding.

[11] Yang, H. (2019) Bert Meets Chinese Word Segmentation.

[12] Meng, Y., Wu, W., Wang, F., Li, X., Nie, P., Yin, F., Li, M., Han, Q., Sun, X. and Li, J. (2019) Glyce: Glyph-Vectors for Chinese Character Representations. *Proceedings of the* 33*rd International Conference on Neural Information Processing Systems*, Vancouver, 8-14 December 2019, 2746-2757.

[13] Peng, M., Ma, R., Zhang, Q. and Huang, X. (2019) Simplify the Usage of Lexicon in Chinese NER. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Florence, 28 July-2 August 2019, 5951-5960.

[14] Liu, W., Fu, X., Zhang, Y. and Xiao, W. (2021) Lexicon Enhanced Chinese Sequence Labeling Using BERT Adapter. *Proceedings of the* 59*th Annual Meeting of the Association for Computational Linguistics and the* 11*th International Joint Conference on Natural Language Processing*, 1-6 August 2021, 5847-5858. https://doi.org/10.18653/v1/2021.acl-long.454

[15] Zhang, Z., Zhang, H., Chen, K., Guo, Y., Hua, J., Wang, Y. and Zhou, M. (2021) Mengzi: Towards Lightweight yet Ingenious Pre-Trained Models for Chinese.

[16] Lafferty, J.D., McCallum, A. and Pereira, F. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the International Conference on Machine Learning*, Williamstown, 28 June-1 July 2001, 282-289.

[17] 彭佳元. 融合字形特征的中文命名实体识别方法研究[D]: [硕士学位论文]. 上海: 上海交通大学, 2019. https://doi.org/10.27307/d.cnki.gsjtu.2019.004104

[18] Xiao, L. and Pennington, J. (2022) Synergy and Symmetry in Deep Learning: Interactions between the Data, Model, and Inference Algorithm. *Proceedings of the* 39*th International Conference on Machine Learning*, Baltimore, 17-23 July 2022, 24347-24369.

[19] Song, Y., Shi, S., Li, J. and Zhang, H. (2018) Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. *Proceedings of the* 2018 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, New Orleans, 1-6 June 2018, 175-180. https://doi.org/10.18653/v1/N18-2028

[20] Sun, M., Chen, X., Zhang, K., Guo, Z. and Liu, Z. (2016) THULAC: An Efficient Lexical Analyzer for Chinese. https://github.com/thunlp/thulac

[21] Li, B., Lu, Y., Pang, W., *et al*. (2023) Image Colorization Using CycleGAN with Semantic and Spatial Rationality. *Multimedia Tools and Applications*, **82**, 21641-21655. https://doi.org/10.1007/s11042-023-14675-9

[22] Peng, N. and Dredze, M. (2015) Named Entity Recognition for Chinese Social Media with Jointly Trained Embeddings. *Proceedings of the* 2015 *Conference on Empirical Methods in Natural Language Processing*, Lisbon, 17-21 September 2015, 548-554. https://doi.org/10.18653/v1/D15-1064

[23] Zhang, Y. and Yang, J. (2018) Chinese NER Using Lattice LSTM. *Proceedings of the* 56*th Annual Meeting of the Association for Computational Linguistics*, Melbourne, 15-20 July 2018, 1554-1564. https://doi.org/10.18653/v1/P18-1144

[24] Levow, G.-A. (2006) The Third International Chinese Language Processing Bakeoff: Word Segmentation and Named Entity Recognition. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, Sydney, 22-23 July 2006, 108-117.

[25] Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M. and Liu, Q. (2019) ERNIE: Enhanced Language Representation with Informative Entities. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, 28 July-2 August 2019, 1441-1451. https://doi.org/10.18653/v1/P19-1139

[26] Sun, Z., Li, X., Sun, X., Meng, Y., Ao, X., He, Q., Wu, F. and Li, J. (2021) ChineseBERT: Chinese Pretraining Enhanced by Glyph and Pinyin Information. *Proceedings of the* 59*th Annual Meeting of the Association for Computational Linguistics and the* 11*th International Joint Conference on Natural Language Processing*, 1-6 August 2021, 2065-2075. https://doi.org/10.18653/v1/2021.acl-long.161

[27] Li, S., He, W., Shi, Y., Jiang, W., Liang, H., Jiang, Y., Zhang, Y., Lyu, Y. and Zhu, Y. (2019) DuIE: A Large-Scale Chinese Dataset for Information Extraction. *Proceedings of the Natural Language Processing and Chinese Computing*, Dunhuang, 9-14 October 2019, 791-800. https://doi.org/10.1007/978-3-030-32236-6_72

[28] Xu, J., Wen, J., Sun, X. and Su, Q. (2017) A Discourse-Level Named Entity Recognition and Relation Extraction Dataset for Chinese Literature Text.

[29] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.D.L., Hendricks, L.A., Welbl, J., Clark, A., *et al*. (2022) Training Compute-Optimal Large Language Models.