计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241200035

基于融合模型的警情地址相似度计算

张 硕 季 铎
中国刑事警察学院公安信息技术与情报学院 沈阳 110000
(2926637201@qq.com)

摘 要 随着大数据技术在公安领域的广泛应用,警情响应速度的提升已成为推动公安现代化及高效运作的核心目标之一。警情快速响应系统通过自动派警机制替代传统人工派警,其核心依赖于模型对警情地址的精准识别。然而,警情地址与普通地址在特征表现上存在显著差异,现有的商业化地址匹配模型在处理警情地址时,常常存在适配性不足的问题。为解决这一问题,提出了一种结合地址分级和拼音信息的改进方法,旨在替代传统深度学习算法,以应对商业化地址计算模型在警情地址识别中的局限性。该方法针对中文警情地址中的特殊词组、多层次地址结构、同音异字及错别字等特点进行优化。通过预训练模型、数据增强、地址分级及拼音信息编码等技术手段,研究构建并训练一种专用于警情地址相似度计算的高效模型,显著提高中文警情地址的识别准确性与适配能力。

关键词:警情地址;地址分级;拼音;深度学习;预训练;数据增强
中图分类号 TP391

Calculation of Police Incident Address Similarity Based on Fusion Model

ZHANG Shuo and JI Duo
School of Public Security Technology and Information,Criminal Investigation Police University of China,Shenyang 110000,China

Abstract With the widespread application of big data technology in the field of public security,the improvement of police response speed has become one of the core goals to promote the modernization and efficient operation of public security. The rapid response system for police incidents replaces traditional manual dispatch with an automatic dispatch mechanism,and its core relies on the model's accurate identification of police addresses. However,there are significant differences in feature representation between police addresses and regular addresses,and existing commercial address matching models often suffer from insufficient adaptability when dealing with police addresses. To address this issue,this paper proposes an improved method that combines address grading and pinyin information,aiming to replace traditional deep learning algorithms and address the limitations of commercial address calculation models in police address recognition. This method is optimized for the special phrases,multi-level address structure,homophones,and misspellings in Chinese police addresses. By using techniques such as pre-training models,data augmentation,address grading,and Pinyin information encoding,this paper aims to develop and train an efficient model specifically designed for calculating the similarity of police addresses,significantly improving the recognition accuracy and adaptability of Chinese police addresses.

Keywords Police address,Address classification,Pinyin,Deep learning,Pre-training,Data augmentation

1 引言

近年来,随着大数据技术在公安领域应用需求的不断增长,警情快速响应的重要性日益凸显。尤其是在利用大数据技术高效处理警情的背景下,自动派警系统逐渐取代传统的人工派警方式,成为提升警情处置速度、效率和精确度的重要手段。自动派警系统的核心在于,系统能否准确识别警情地址。因此,构建有效的模型并将其应用于大数据驱动的自动派警系统中,对于提高警情响应速度具有重要意义。然而,由于人们描述地址的方式多样,且常常存在错别字和口语化表达等问题,这为警情地址的精准识别带来挑战。此外,中文地址的相似度计算与普通文本的相似度计算有所不同,无法简单地套用传统的文本相似度算法。地址相似度和文本相似度的定义规则及词向量标准存在显著差异。中文地址中存在一些特定的地名、街道名或商铺名,这些词汇往往只是单纯的标识符,并未包含深层语义联系。例如,"教育路"中的教育与"教学路"中的教学,在地址相似度模型中应被视作两个差异较大的词汇,但在基于普通文本的预训练相似度模型中,这两个词汇的向量表示可能非常接近。因此,直接使用传统中文文本相似度算法进行中文地址相似度计算并不可行。然而,中文地址中也存在具有深层语义联系的词组,如"儿童急诊室"和"儿科"指代同一地方。理想的地址相似度计算模型应能够同时判断字面上的词语相似性,并能够识别具有深层语义联系的词组。

基金项目:辽宁网络安全执法协同创新中心资助
This work was supported by the Liaoning Collaboration Innovation Center For CSLE.
通信作者:季铎(18640037173@163.com)

为应对上述问题，本文提出一种融合模型，结合数据集标准化处理和深度学习技术，旨在训练出一个专门用于计算中文警情地址相似度的模型。该模型能够有效解决中文地址中的特殊词组、多层次地址格式及同音错别字等问题，从而更好地服务于公安警情的高效处理。基于预训练深度学习模型，本文进一步引入地址分级和融合中文地址及其汉语拼音的输入方式，旨在提升模型在警情地址相似度计算中的适应性。通过这种创新的输入方式，模型能够更准确地处理警情地址的复杂性，提升识别和匹配的精度。

2 地址相似度计算的相关研究

地址相似度计算的研究可以追溯到 20 世纪 90 年代和 21 世纪初。随着信息检索和数据库管理技术的发展，研究人员开始关注如何有效地处理和比较文本数据。这一领域的研究可以大致分为以下几个主要方向。

2.1 字符串相似度算法

此类算法主要基于编辑距离，即通过计算将一个字符串转换为另一个字符串所需的最小编辑操作数来评估相似度。典型的算法包括 Levenshtein 距离，Jaccard 相似系数，Cosine 相似度，以及 Jaro-Winkler 距离。这些方法[1-5]能有效处理字符级的匹配问题，但对于处理地址中的缩写、拼写错误以及语义上的变化则显得不足。Diao 等[6]通过论述拼音的作用验证拼音在地址相似度计算中的可行性。

2.2 基于领域知识的计算方法

在领域知识的帮助下，地址匹配可以更加精准。例如，地理编码技术，Lin 等[7]通过将地址转换为地理坐标，通过坐标的接近程度来判断地址的相似性。此外，将地址进行分块和分词处理，如将街道名、门牌号、城市和邮编等独立考虑，并通过加权方法组合各部分的相似度，也是一种有效的匹配策略。Yu 等[8]通过领域特定的词典和规则来处理中文地址的相似度。

2.3 自然语言处理技术和知识图谱

随着自然语言处理技术的发展，语义分析方法在地址相似度计算中的应用日益增多。这类方法通过将地址中的词语转换为向量，如通过 Word2Vec 或 GloVe[9-11]模型计算向量间的余弦相似度，以此捕捉词语间的潜在语义关系。另外，通过句子向量模型[12-13]，如 BERT 或 Sentence-BERT 能够提供更全面的地址向量表示，进而通过向量相似度来评估地址间的相似度。近年来，深度学习方法在地址相似度计算中得到了广泛应用。首先，Zhou[14]在 Transformer 模型的基础上引入地理实体类别嵌入和实体级自注意力机制，显著增强深度学习模型的学习能力。接着，Zhang[15]通过预训练模型（如 BERT 和 GPT）生成上下文感知的向量表示，为计算相似度提供新的技术手段。在此基础上，Yang[16]采用双向长短期记忆网络（Bi-LSTM）混合结构，处理复杂的地址结构并提高匹配准确性，验证 Bi-LSTM 的有效性。随后，Peng 等[17]进一步使用 Bi-LSTM-CRF 模型，结合自注意力机制和序列生成网络，达到了较为理想的效果。通过这些研究，深度学习在地址相似度计算中的表现得到持续优化。此外，Chen[18]探讨了知识图谱在中文地址中的应用潜力，验证其可行性，提出了新的方法论。

3 地址标准化

3.1 警情地址实例及其特点

在许多情况下，警情的突发性导致报警地址呈现出较强的随意性和口语化特点。此外，由于报警人在紧张情绪的影响下，往往会产生冗余或过于详细的描述；而在时间紧迫的情况下，报警地址的描述可能会显得过于简略。同时，为提高响应效率，目前广泛采用语音转文字技术记录报警信息。然而，这一技术的应用也带来了新的问题，尤其是由语音识别引起的错别字问题愈加显著。这种现象不仅影响信息传递的准确性，还可能在紧急情况下延误救援行动。表 1 列出了普通地址与一些常见警情地址在描述特点上的差异。

表 1 警情地址特点类型

Table 1 Characteristics and types of police addresses

普通地址	警情地址	警情地址特点
和平区胜利大街雅致菜市场	和平区胜利大街雅致菜市场一楼进门右拐第 3 个蔬菜摊位	口语化严重的警情地址
和平区北市场街道西塔文化步行街	西塔步行街	表述简略的警情地址
和平区文安路小核桃烤肉	和平区文安路削核桃烤肉	语音转写造成错别字的警情地址，将“小”转写为“削”

与普通地址相比，警情地址在市、区、街道级别的标准地址部分差异较小。例如，表 1 中普通地址和警情地址中的“和平区胜利大街雅致菜市场”完全一致。然而，在具体的详细地址部分，警情地址通常采用口语化的描述方式，例如表 1 中警情地址的“一楼进门右拐第三个蔬菜摊位”。这种口语化的表述方式导致警情地址的前后部分特征存在显著的不一致性，这也是警情地址的一个重要特点。

3.2 国家标准的地址标准化

地理信息系统（GIS）、计算机科学和语言学的研究表明，中文地址体系具有明显的层次性结构。中国的行政区划主要划分为四级：省级行政区、地级行政区、县级行政区和乡级行政区。省级行政区包括 23 个省、5 个自治区、4 个直辖市和 2 个特别行政区，共计 34 个。例如，河北省、山西省、辽宁省等为省份；内蒙古自治区、广西壮族自治区等为自治区；北京市、上海市、天津市、重庆市为直辖市；香港特别行政区和澳门特别行政区为特别行政区。地级行政区包括 293 个地级市、7 个地区、30 个自治州和 3 个盟，共计 333 个。地级行政区进一步细分为市辖区、县级市、县、自治县等。县级行政区包括 977 个市辖区、394 个县级市、1 301 个县、117 个自治县、49 个旗、3 个自治旗、1 个特区和 1 个林区，共计 2843 个。县级行政区下设乡、镇、街道等。乡级行政区包括 8984 个街道、21 389 个镇、7 116 个乡、153 个苏木、966 个民族乡、1 个民族苏木和 2 个县辖区，共计 38 602 个。该层次化结构不仅有助于实现精确的地理定位，同时，相关政府部门也会定期发布详细的行政区划数据库，进一步支持各类地理信息应用。

3.3 地址分级

地址分级是将地址数据按照不同的行政层级进行拆分，

具体层级包括：1）区；2）街道；3）社区；4）街；5）路；6）具体位置。首先，构建市、区、社区、街道等层级的集合词典，并基于这些词典构建本文所提出的地址标准化方法。在此基础上，进行地址的分级处理。对地址进行分级的伪代码示例如算法 1 所示。

算法 1　地址分级代码

输入：完整地址

输出：区地址，社区地址，街道地址，街地址，路地址，具体地址

1. 建立词典

FUNCTION extract_components(address，districts，streets，communities)

2. 匹配区

District← find_best_match(address，districts)

3. 匹配社区

Community← find_best_match(address，communities)

4. 匹配街道

street← find_best_match(address，streets)

5. 正则表达式匹配街

street_part，address ← split_street(address)

6. 正则表达式匹配路

road_part，remaining_text ←split_road(address)

7. 输出结果

RETURN[district，street，community，street_part，road_part，remaining_text]

在数据处理过程中，前三级地址通过预先构建的词典进行匹配。对于第四和第五级地址信息，可以采用正则表达式进行精确匹配。完成前五级地址的匹配后，剩余部分的数据即代表第六级具体位置。地址分级的处理流程如图 1 所示。

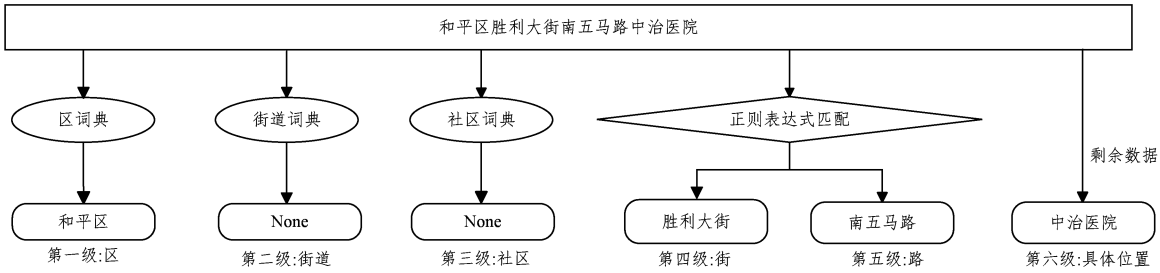

图 1　地址分级流程

Fig. 1　Address grading process

4　基于融合模型的地址相似度计算方法

4.1　Chinese-RoBerta-wwm-ext 文本计算模型

Chinese-RoBerta-wwm-ext 是一种基于词嵌入的预训练模型，由哈尔滨工业大学讯飞联合实验室发布。该模型不仅支持中文数据的学习，还能处理英文数据，符合中文地址与拼音结合的数据输入需求。模型基于多头注意力机制，已被证明在中文文本分类任务中表现出色。术语“wwm”代表“word-level whole word masking”，即在预训练过程中应用整词掩码（Whole Word Masking，WWM）技术。对于中文这种缺乏空格分隔的语言，整词掩码策略尤为关键，有助于模型在处理连续词汇时更准确地捕捉词边界信息。Chinese-RoBerta-wwm-ext 在预训练阶段利用大规模的中文语料库，能够捕捉丰富的语言特征和上下文信息。整词掩码策略使得模型能够更准确地理解词汇边界，从而在诸如命名实体识别、情感分析、问答系统等下游任务中显著提升准确性与鲁棒性。此外，Chinese-RoBerta-wwm-ext 融入“扩展词嵌入”（Extended Word Embeddings）概念，该方法不仅关注单个词汇，还充分考虑词汇之间的相互关系。这一扩展的词嵌入策略大大增强了模型在处理复杂语义关系方面的能力。该模型基于 Transformer 架构，利用自注意力机制处理输入序列，自注意力机制通过多个“头”并行处理序列的不同方面，从而增强模型的表达能力。

$$MultiHead(Q,K,V)=Concat(head_1,\cdots,head_h)W^o \quad (1)$$

$$\text{where } head_i=Attention(QW_i^Q,KW_i^K,VW_i^V) \quad (2)$$

$$Attention(Q,K,V)=\text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \quad (3)$$

Chinese-RoBerta-wwm-ext 使用的整词掩码技术特别适用于中文地址的相似度计算。传统的掩码技术仅对单个字符进行向量化处理，这使得模型难以捕捉到地址中某些词汇的深层次含义，尤其是对于那些没有明确字面意义但在地址表达中起重要作用的词汇。例如，某些词汇仅作为地名或街道名称存在，而普通掩码方法无法有效识别这些词汇的语义关系。采用整词掩码技术后，模型能够将整个词作为一个单位进行处理，从而更准确地学习到该词的整体语义和相关特征，提升了对中文地址中具有深层次意义词汇的理解能力。图 2 给出了整词掩码技术在地址处理中的优势。

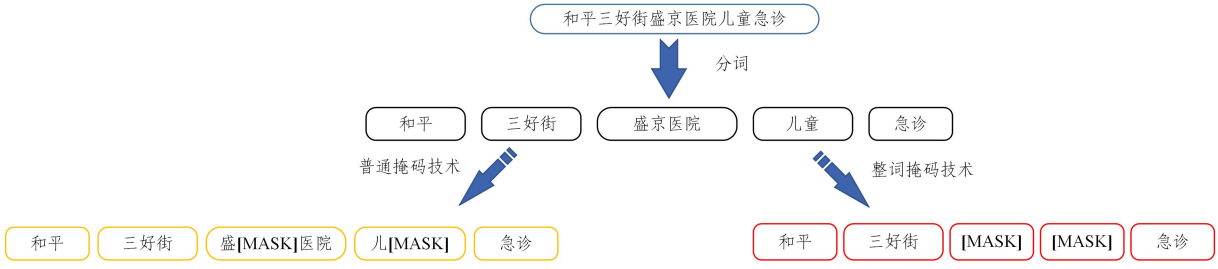

图 2　整词掩码技术和普通掩码技术区别

Fig. 2　Difference between whole word masking technology and ordinary masking technology

4.2 BiLSTM 文本计算方法

BiLSTM(双向长短期记忆网络)是一种特殊类型的循环神经网络(RNN),在序列数据处理任务中表现出色。与传统的 LSTM 网络仅能按单向时间序列处理信息不同,BiLSTM 通过结合两个独立的 LSTM 网络,能分别处理时间序列的正向和反向信息,使得网络能够同时获取过去和未来的上下文信息,如图 3 所示。该结构特别适用于需要考虑全局序列信息的任务。

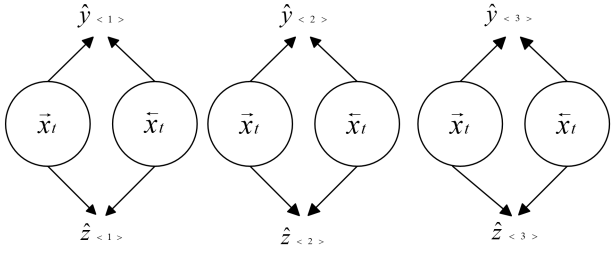

图 3 BiLSTM 结构

Fig. 3 Structure of BiLSTM

时间 t 的输出预测值为:

$$\hat{y}_{\langle t\rangle}=g(W_y\cdot[\vec{x}_{\langle t\rangle};\overleftarrow{x}_{\langle t\rangle}]+b_y) \tag{4}$$

在 BiLSTM 中,每个 LSTM 单元通过以下关键方程来更新其状态。

Step1 遗忘门:控制前一个隐藏状态h_{t-1}应该保留多少到当前状态。

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f) \tag{5}$$

Step2 输入门:决定有多少新的输入信息x_t应被引入到单元状态。

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i) \tag{6}$$

Step3 单元状态更新:根据输入和过去的单元状态C_{t-1}创建候选列表。

$$\widetilde{C}_t=\tanh(W_c\cdot[h_{t-1},x_t]+b_c) \tag{7}$$

Step4 最终单元状态:当前时刻的单元状态C_t是前一时刻状态的遗忘和当前候选状态的加权和。

$$C_t=f_t*C_{t-1}+i_t*\widetilde{C}_t \tag{8}$$

Step5 输出门:o_t决定当前单元状态的哪些部分将输出。

$$o_t=\sigma(W_o\cdot[h_{t-1},x_t]+b_o) \tag{9}$$

Step6 隐藏状态更新:最终的输出隐藏状态h_t是输出门和单元状态的激活值的乘积。

$$h_t=o_t*\tanh C_t \tag{10}$$

4.3 卷积神经网络文本计算方法

卷积神经网络(CNN)在文本处理中广泛应用于自动提取句子或文档中的关键特征,这些特征对于理解文本语义至关重要。与传统的自然语言处理(NLP)模型相比,CNN 通过卷积层能够有效捕捉局部特征,从而在分类、情感分析或主题识别等任务中提供更加精确的结果。在中文地址处理过程中,首先通过 RoBERTa 模型将地址转换为数值形式的词向量,随后 CNN 对这些向量进行处理。通过使用多个不同大小的卷积滤波器,CNN 能够提取文本中的 n-gram 特征。接着,池化层对卷积层生成的特征图进行降维处理,从而简化特征表示,并最终输出分类结果。

文本卷积操作如下:

$$c_i=f(W\cdot h_{i,i+h-1}+b) \tag{11}$$

4.4 自注意力机制

Attention 机制最初被提出用于解决传统循环神经网络(RNN)和长短期记忆网络(LSTM)在处理长序列时,信息随着时间推移逐渐稀释的问题。Attention 机制允许模型在生成输出时,动态地聚焦输入序列中与当前任务相关的部分,而不是盲目依赖于固定区域的信息。这一机制显著提升了模型对长序列信息的处理能力,尤其是在需要理解和处理复杂关系的任务中。具体而言,Attention 机制首先将输入数据转换为查询(Q)、键(K)和值(V)。然后,通过计算查询向量 $\boldsymbol{Q}$ 和键向量 $\boldsymbol{K}$ 的点积,得到注意力分数(Scores)。接着,利用 Softmax 函数对这些注意力分数进行归一化处理,使得分数的和为 1。归一化后的分数反映每个值(V)在最终输出中的相对重要性,最后,模型通过对所有值(V)进行加权求和,计算最终的输出。这种机制使得模型能够根据任务需求灵活调整信息的关注重点,提高模型在复杂任务中的表现能力。

Step1 注意力分数:

$$Scores=\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}} \tag{12}$$

Step2 Value 的相对重要性:

$$AttentionWeights=\mathrm{Softmax}(Scores) \tag{13}$$

Step3 加权求和:

$$Output=Attention\ Weights\cdot V \tag{14}$$

4.5 地址相似度计算方法实验设计

在中文长文本的比较中,常见的深度学习方法如 LSTM 和自注意力机制,通常用于捕捉前后文的上下文特征。然而,在警情地址的处理过程中,前后文本特征的显著不一致性使得传统的深度学习方法难以有效应用。具体而言,前五级地址(如省、市、区、街道、路名)可视为 5 个集合的随机组合,主要由字面含义明确的词组构成,如区县名、路名、街道名等。这些地址级别通常具有较高的规范性和一致性。与之相比,第六级地址则呈现出高度的随机性和非结构化特征,既包含深层次语义的词组(如特定店名、功能性名称等),又包含单纯的字面描述。因此,从理论上讲,采用地址分级策略对地址进行处理,能够更有效地帮助模型学习到不同级别地址的特征,从而提升模型在警情地址识别中的表现。地址分级的策略示意图如图 4 所示。

在中文地址的相似度计算中,错别字及相似描述的地址区分是一个显著的难点。具体而言,同音错别字和同一地址的不同书写形式可能指向相同地点。例如,“仲景药房”和“中景药房”仅存在一字之差,但实际上指代的是同一地址,因此该对地址的标签应为“1”。而一些方位词的差异,如“东边”和“西边”,虽然仅一字之差,却可能指向完全不同的地址,因此其标签应为“0”。在没有拼音信息的情况下,数据集中存在大量仅一字之差的地址对,其中一些是相似的(标签为“1”),而另一些则不相似(标签为“0”)。这些细微的差异使得传统深度学习模型在处理这些地址对时面临较大挑战。由于错别字通常具有较高的发音相似度且拼音差异较小,而完全不同的字则在拼音上存在较大差异,引入拼音信息,理论上可以帮助模型更有效地区分这些细微的语音特征,从而提高对地址相似度的判别能力。

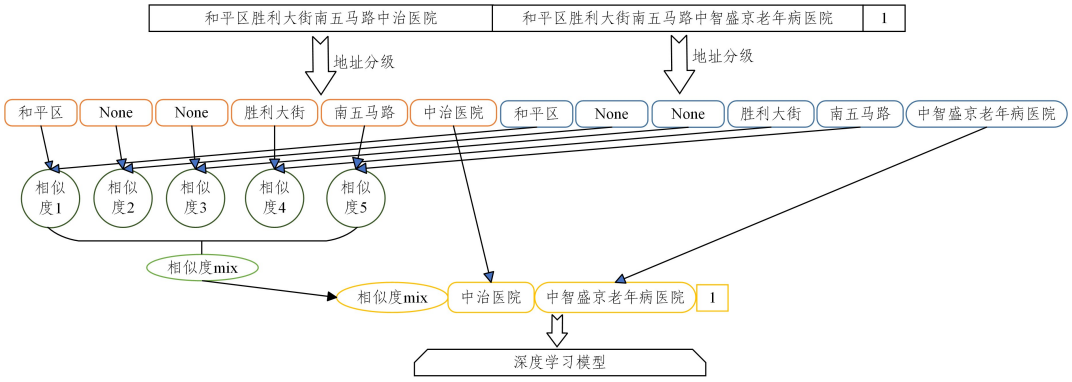

图 4　地址分级输入模型策略

Fig. 4　Address grading input model strategy

为探讨地址分级与拼音信息的引入是否优于传统深度学习算法在处理警情地址时的效果，本文设计了 3 组实验，具体如下。

(1)地址分级与前后文特征学习算法对比实验：该实验旨在比较引入地址分级的训练方法与采用能够学习文本前后文特征的深度学习算法(如 LSTM，Transformer 等)的训练方法在处理警情地址相似度计算中的表现差异。

(2)加入拼音与未加入拼音的对比实验：本实验通过比较加入拼音信息与未加入拼音信息的中文地址在相似度计算中的表现，分析拼音信息对提升模型区分错别字及相似地址能力的影响。

(3)地址分级结合拼音与前后文特征学习算法对比实验：该实验结合地址分级与拼音信息，同时与能学习前后文特征的深度学习算法进行比较，旨在评估两者结合使用时在警情地址相似度计算中的综合效果。

本文共运行 16 个模型，涵盖 4 种不同的深度学习模型与 4 种数据处理方式的组合。具体而言，实验中采用的 4 类模型分别为：Chinese-RoBerta-wwm-ext 模型、BiLSTM-Chinese-RoBerta-wwm-ext 模型、BiLSTM-Chinese-RoBerta-wwm-ext-CNN 模型，以及 BiLSTM-Chinese-RoBerta-wwm-ext-Attention 模型。与此同时，4 种数据处理方式包括：使用原始数据集、采用地址分级策略、融合拼音信息的策略，以及结合地址分级和拼音信息的策略。通过这些实验，旨在验证地址分级、BiLSTM 模型以及拼音引入在中文警情地址相似度计算中的效果和优势。

实验过程如表 2 所列。

表 2　对比实验过程

Table 2　Comparative experimental process

实验	对比模型
实验一	C-RoBerta，BiLSTM-RoBerta，BiLSTM-RoBerta-CNN，BiLSTM-RoBerta-Attention
实验二	C-RoBerta，BiLSTM-RoBerta，BiLSTM-RoBerta-CNN，BiLSTM-RoBerta-Attention，PC-RoBerta，P-BiLSTM-RoBerta，P-BiLSTM-RoBerta-CNN，P-BiLSTM-RoBerta-Attention
实验三	PC-RoBerta，BiLSTM-RoBerta，BiLSTM-RoBerta-CNN，BiLSTM-RoBerta-Attention

注：P：引入拼音信息的策略；C：采用地址分级策略；RoBerta：原模型 Chinese-RoBerta-wwm-ext。

通过上述 3 组实验，本研究旨在深入探讨不同方法在中文警情地址相似度计算中的适用性与局限性，以期找到更为有效的解决方案，从而提升警情响应的准确性与效率。第一组实验将比较地址分级方法与深度学习算法在地址相似度计算中的表现，以评估哪种方法更适合处理警情地址中的复杂情况。第二组实验将验证拼音信息的引入能否有效改善模型的学习效果，特别是在处理错别字和同音词等问题时的表现。第三组实验则重点考察地址分级与拼音结合的方式能否弥补当前深度学习模型在警情地址处理中的不足，提升模型的泛化能力和准确性。

最终，基于上述实验结果，本文构建了一个综合模型，该模型不仅能够精准计算地址间的相似度，还具备地址纠错和地址分类等附加功能。该综合模型的提出，将为公安领域的相关工作提供强有力的技术支持，显著提高警情处理的效率和准确性。

5　实验对比

5.1　数据集

本文所使用的数据集均来源于公安部门的警情记录。在这些记录中，存在同一警情多次报警的情况，比例超过 30%。因此，从中提取所有相关的地址信息，并进行数据清洗。为确保数据的质量和具有代表性，采用人工标注的方式对数据进行筛选和修正。

在定义地址相似性时，本文遵循以下原则：在处理警情记录时，若两个地址描述指向同一实际位置，则将其视为正样本；而对于表示不同地址的描述，则视为负样本。此外，本文根据原始数据手动构造一些示例，旨在突出模型需要学习的关键特征，尤其是细节特征，以帮助模型更有效地识别和学习这些特征。为增强模型的泛化能力，在数据处理阶段，本文引入数据增强技术。通过对训练数据进行随机扰动，例如替换、插入或删除地址中的部分内容，生成更多样本，从而有效避免模型对特定地址格式的过拟合问题。该数据集不仅包括准确的地址信息，还涵盖各种形式的错误地址和变体地址，以便模型能够有效学习处理这些特殊情况，从而提高其在实际应用中的鲁棒性和准确性。与此同时，本文还采用如 dropout 和权重衰减等正则化技术，以防止模型在训练过程中对某些特征产生过度依赖，进一步提升了模型的鲁棒性和泛化能力。

以下情形被视为相同地址：

(1)在原始数据集中，针对同一警情地址的不同描述。

(2)描述中有所省略，但仍指向相同地点的地址，例如省略具体的街道或社区信息。

(3)存在错别字的地址描述，若该描述仍能指向同一实际

位置。

(4)名称存在细微差异,但实际指向相同地点的情况,例如“如家宾馆”与“如家酒店”。

以下情形则被视为不同地址:

(1)原始数据集中,针对不同警情地址的描述。

(2)名称相同但地理位置不同的地址,例如位于不同区域的“蜜雪冰城”。

(3)名称一致但地址类型不同的情况,例如“仲景药房”与“仲景宾馆”。

数据集样例如表 3 所列。

表 3 数据集样例

Table 3 Sample of dataset

Sentence1	Sentence2	label
和平区胜利大街南五马路中治医院院内	和平区胜利大街南五马路中智盛京老年病医院 4 楼	1
和平区三好街云顶大厦附近	和平区三好街云顶大厦旁边	1
铁西区建设东路	大东区滂江街家乐福超市停车场	0
皇姑区塔湾街地铁站 A 出口	皇姑区塔湾街地铁站 B 出口	0
和平区长白二街麒麟酒吧	和平区长白二街麒麟酒馆	1
……		

5.2 实验评价指标

实验采用准确率和 F1 值来评估实验过程。

(1) Accuracy(准确率):是指模型预测正确的样本数占总样本数的比例。

$$准确率=\frac{正确预测的样本数}{总样本数}$$

(2)Precision(精确率):测量模型在所有预测为正类的样本中,真正为正类的比例。

$$精确率=\frac{真正类}{真正类+假正类}$$

(3)Recall(召回率):测量模型在所有真正的正类样本中,预测为正类的比例。

$$召回率=\frac{真正类}{真正类+假负类}$$

(4) F1 值(F1 Score):F1 值是精确率和召回率的调和平均数,是精确率和召回率的综合评价指标。

$$F1\ 值=2\times\frac{精确率\times召回率}{精确率+召回率}$$

5.3 实验结果

总实验结果如图 5 所示。

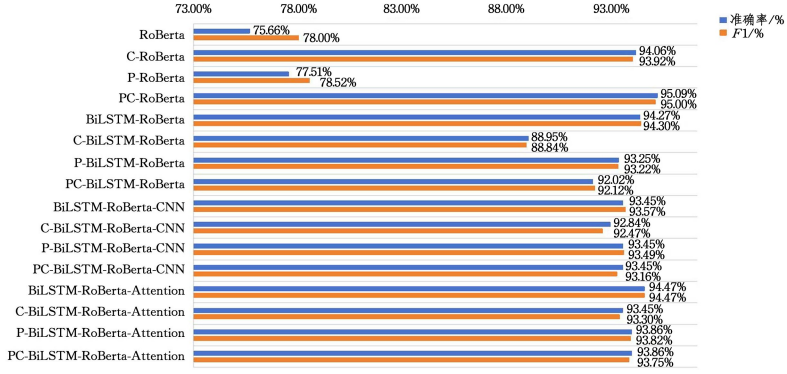

图 5 总实验结果

Fig. 5 Overall experimental results

5.4 实验结果分析

在训练的 16 个模型中,关于地址分级策略的对比实验效果如图 6 所示。加入拼音的对比实验效果则分别如图 7、图 8 所示。最后,结合地址分级与拼音信息,并与前后文特征学习算法进行对比的实验结果如图 9 所示。

从图 6 的实验结果可以看出,使用 Chinese-RoBerta-wwm-ext 模型时,地址分级策略的引入显著提高了模型的表现。具体而言,在未采用地址分级的情况下,模型的准确率和 F1 值仅约为 70%;而在应用地址分级后,这些指标显著提升至 94%以上。进一步的对比实验中,尽管在原模型 Chinese-RoBerta-wwm-ext 的基础上加入 BiLSTM、CNN 和自注意力机制等深度学习算法,模型的准确率和 F1 值有所下降。但结果表明,采用地址分级策略的 Chinese-RoBerta-wwm-ext 模型在准确率和 F1 值上,几乎与未采用地址分级的 BiLSTM、CNN 或自注意力机制模型的表现相当。这一结果表明,地址分级与前后文特征学习能力较强的深度学习算法在警情地址相似度计算任务中的效果相当。

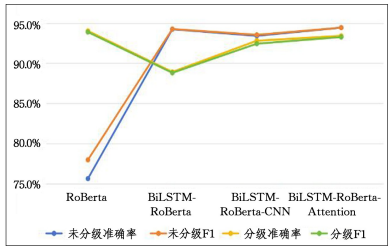

图 6 地址分级实验结果

Fig. 6 Experimental results of address grading

从图 7 的实验结果可以看出,在引入拼音信息后,采用地址分级策略的 Chinese-RoBerta-wwm-ext 模型的表现得到显著提升,并且超过结合 BiLSTM、CNN 和自注意力机制的深度学习模型的效果。然而,当在原始模型 Chinese-RoBerta-wwm-ext 的基础上加入 BiLSTM、CNN 或自注意力机制,并且未采用地址分级时,加入拼音后模型的表现却出现下降。为进一步分析,还对比了在 BiLSTM、CNN 和自注意力机制模型中加入地址分级并引入拼音后的实验结果,如图 8 所示。

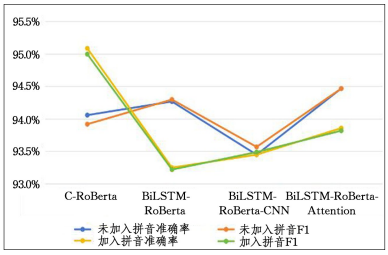

图 7　融合拼音信息实验结果

Fig. 7　Experimental results of integrating Pinyin information

通过图 8 的对比实验结果进一步表明，在采用地址分级策略的同时，结合能够学习前后文特征的 BiLSTM、CNN 和自注意力机制时，拼音的引入显著提升了模型的整体表现。

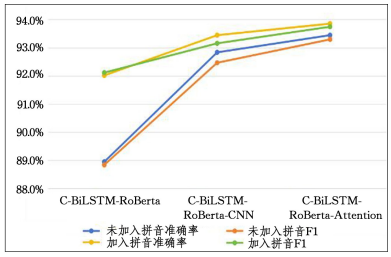

图 8　融合拼音对比实验结果

Fig. 8　Comparison experiment results of fusion Pinyin

与图 8 中未采用地址分级的模型结果相比，拼音在地址分级的情况下表现出更为显著的优势。尤其在地址分级的背景下，其作用更加突出。相比之下，由于错别字在文本中所占比例较小，单纯加入拼音可能会引入冗余特征，反而可能对模型性能产生负面影响。然而，在地址分级策略下，由于每一层级的地址文本较为简短，拼音差异更容易被放大，从而便于模型更好地区分细微差异。因此，拼音在地址分级的情况下更能发挥其优势，有助于提升相似度计算的精度和鲁棒性。

通过对 16 个模型结果的对比分析，结果显示 4 类模型中表现最优的是采用地址分级并引入拼音的 Chinese-RoBerta-wwm-ext 模型，如图 9 所示。其效果优于所有基于学习前后文特征的深度学习模型。这一结果表明，结合地址分级和拼音的策略在警情地址相似度计算中展现出显著优势，超越了传统深度学习方法中依赖前后文特征的算法。该策略不仅有效提升了模型的准确率和 F1 值，还为警情响应提供更为高效和精确的支持。

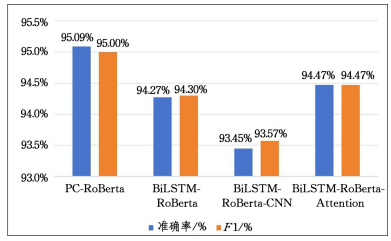

图 9　4 类模型最佳性能对比

Fig. 9　Comparison of the best performance of four types of experimental models

通过上述实验，本文最终训练出最适用于计算中文警情地址相似度的模型 PC-Chinese-RoBerta-wwm-ext。该模型是在预训练文本计算模型的基础上，结合地址分级和拼音引入策略，经过数据增强处理后的警情地址数据进行微调而形成的。

整个模型的流程如图 10 所示。

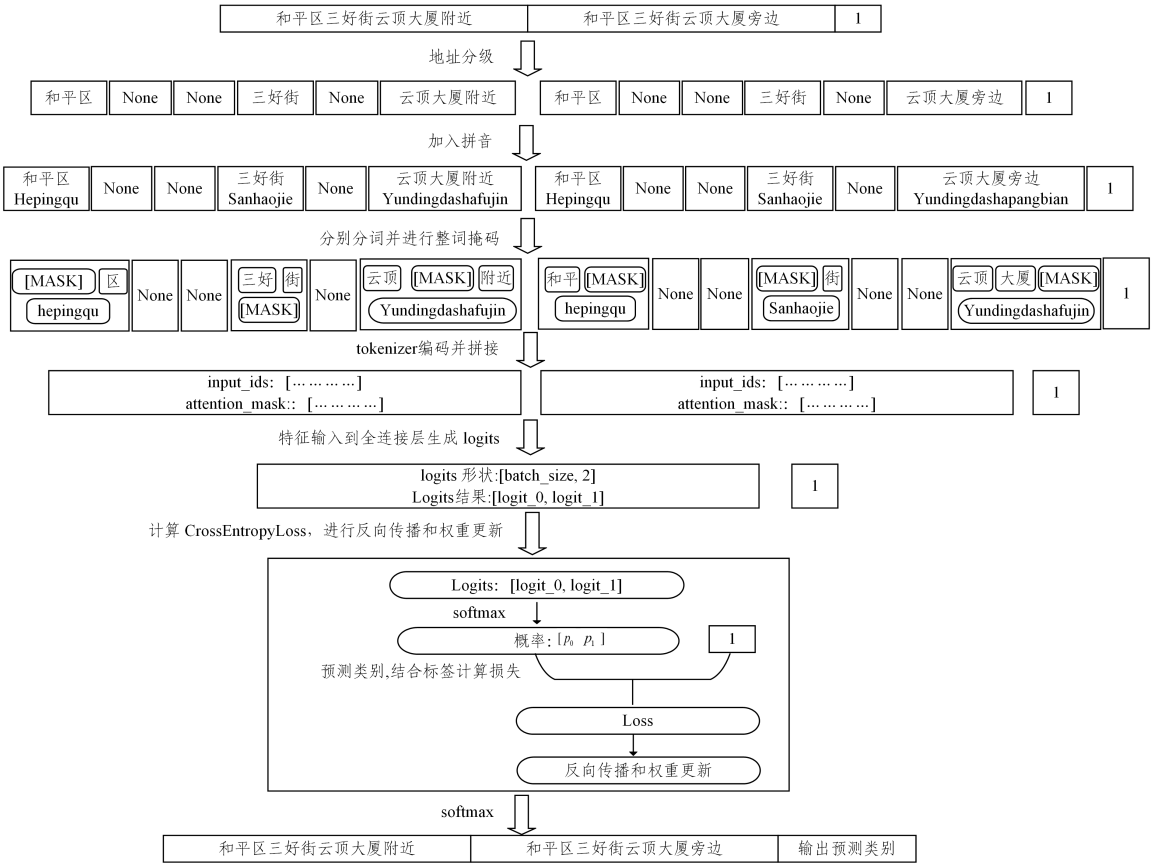

图 10　融合模型流程图

Fig. 10　Flowchart of fusion model

在训练深度学习模型时，同时向模型提供地址的拼音和中文地址。此策略不仅有效解决了地址中错别字的问题，尤其是同音错别字的情况，还处理了地址中同一地点的不同书写形式。由于错别字多为同音词，拼音信息的引入帮助深度学习模型更准确地识别错误字形和相同地址的不同表达方式，进而区分那些表面相似但实际指向不同地址的文本。此外，将拼音与中文地址一同输入模型，还显著提升了模型对地名的泛化能力，使其能够更精确地识别和区分细微差异。引入拼音信息不仅增强模型的语义理解，还提供一种新的正则化机制。在训练过程中，通过随机添加或省略拼音，可以有效防止模型对特定输入格式的过拟合，从而提高其在处理真实世界数据时的鲁棒性。这种随机化策略使得模型在面对变形和错误时具有更强的适应性。此外，拼音转换的应用不仅限于地名，还可以扩展到街道、建筑物等地址组成部分，这能进一步增强模型在处理复杂地址信息时的灵活性和准确性。

结束语 地址分级在处理具有特殊结构特征的警情地址时，能够与深度学习模型产生相似的效果，并有效缓解深度学习算法在该领域中存在的过拟合问题以及难以捕捉细节特征的局限性。通过将地址的拼音与中文地址一同输入深度学习模型，不仅能够有效解决地址识别中错别字和同音异字问题，还显著增强了模型对地名的泛化能力与鲁棒性。拼音信息的引入使得模型能够更加准确地区分那些在视觉上相似的地址，从而提高相似度计算的精度。此外，将拼音与中文地址结合，为深度学习模型提供一种高效的解决方案，极大地提升了模型在处理复杂中文地址信息时的表现。随着自然语言处理技术的持续发展，拼音转换与其他技术的融合有望为地址识别和处理领域带来更多创新性的突破。对于警情地址等具有特殊结构的地址，仍有待进一步探索和开发更适合的算法与策略，以提升模型在此类任务中的表现。

参 考 文 献

[1] KANG M, DU Q, WANG M. A new method of Chinese address extraction based on address tree model[J]. Acta Geodaetica et Cartographica Sinica, 2015, 44(1): 99-107.

[2] KANG M J, DU Q Y, WANG M J. Chinese Address Extraction Method Using Address Tree Model [J]. Journal of Surveying and Mapping, 2015, 44(1): 99-107.

[3] LI X F, SONG Z L, CHEN X X, et al. Research and Implementation of Fuzzy Matching for K-Tree Address [J]. Surveying and Mapping Bulletin, 2018(9): 126-129.

[4] SHI M J. Research on intelligent matching of non-standard Chinese addresses [D]. Xuzhou: China University of Mining and Technology, 2020.

[5] WANG S, ZHUANG S, ZUCCON G. Bert-based dense retrievers require interpolation with bm25 for effective passage retrieval[C] // Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval. 2021: 317-324.

[6] DIAO X C, TAN M C, CAO J J. A string similarity calculation method that integrates multiple editing distances [J]. Computer Application Research, 2010, 27(12): 4523-4525.

[7] LIN Y, KANG M, WU Y, et al. A deep learning architecture for semantic address matching [J]. International Journal of Geographical Information Science, 2020, 34(3): 559-576.

[8] YU T, WANG D, CHEN Q. Chinese address matching method based on pseudo semantic similarity model [J]. Surveying and Mapping Bulletin, 2022(3): 101-106.

[9] LI F, LU Y, MAO X, et al. Multi-task deep learning model based on hierarchical relations of address elements for semantic address matching[J]. Neural Computing and Applications, 2022, 34(11): 8919-8931.

[10] RONG X. word2vec parameter learning explained [J]. arXiv: 1411.2738, 2014.

[11] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C] // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP). Stroudsburg, PA: Association for Computational Linguistics, 2014: 1532-1543.

[12] DEVLIN J, CHANG M W, LEE K, et al. BERT: pretraining of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805, 2018.

[13] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[J]. arXiv: 1907.11692, 2019.

[14] ZHOU H. Geographic coding method based on conditional random fields and spatial inference [D]. Zhengzhou: PLA University of Information Engineering, 2015.

[15] ZHANG H. Research on Chinese Address RESOLUTION and Matching Method Based on BERT Pre trained Model [D]. Nanjing: Nanjing Normal University, 2021.

[16] YANG B. Research on Chinese Address Normalization Technology Integrating Attention Mechanism and Sequence Generation Network [D]. Lanzhou: Lanzhou Jiaotong University, 2023.

[17] PENG Y L, HU S S, WU T. Multi strategy Chinese address matching method [J]. Surveying and Mapping Bulletin, 2022(2): 145-148.

[18] CHEN N Y. Research and Implementation of Semantic Address Matching Method under Privacy Protection [D]. Xi'an: Xi'an University of Electronic Science and Technology, 2023.

ZHANG Shuo, born in 2002, postgraduate. His main research interests include natural language processing and text similarity calculation.

JI Duo, born in 1981, master, associate professor. His main research interests include natural language processing and artificial intelligence.