引文格式：郑诗语,邱芹军,谢忠,等. 基于XLNet的多数据源中文地名匹配方法[J]. 地理空间信息, 2024, 22(8): 59-63. doi:10. 3969/j.issn. 1672-4623.2024.08.013

基于XLNet的多数据源中文地名匹配方法

郑诗语[1,2]，邱芹军[1,2,3*]，谢　忠[1,2]，陶留锋[1,2]，李伟杰[1,2]
（1. 中国地质大学（武汉）计算机学院，湖北 武汉 430074；2. 地理信息系统国家地方联合工程实验室，湖北 武汉 430074；
3. 自然资源部城市国土资源监测与仿真重点实验室，广东 深圳 518000）

摘　要：地址作为社会发展中重要的基础性数据资源，已成为城市地理空间数据化建设的重要组成部分。地名匹配旨在比较表示相同真实世界位置的配对字符串。当前地名匹配方法依赖于字符串相似性独立或多种混合相似性度量方法，这些方法无法有效地捕捉长句子上下文信息，不能充分理解地址含义。因此，提出一种基于XLNet算法的地名匹配方法，利用深度神经网络将一对地名分类为匹配或不匹配。该方法利用长程记忆并使用双信息流注意力掩码对事件序列进行重构，以利用其双向信息建立表征。实验结果表明，该方法可解决长地址匹配问题，模型能较好地理解上下文语义信息，优于先前研究的单个相似度量及基于监督机器学习的方法。

关键词：地名匹配；地名实体；XLNet；Softmax；回归模型
中图分类号：P281　　**文献标志码：**B　　**文章编号：**1672-4623（2024）08-0059-05

Chinese Geographical Name Matching Method with Multiple Data Sources Based on XLNet

ZHENG Shiyu[1,2], QIU Qinjun[1,2,3], XIE Zhong[1,2], TAO Liufeng[1,2], LI Weijie[1,2]
(1. School of Computer Science, China University of Geosciences (Wuhan), Wuhan 430074, China;
2. Laboratory of National Joint Engineering for Geo-information System, Wuhan 430074, China;
3. Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Shenzhen 518000, China)

Abstract: Address, as an important fundamental data resource in social development, has become an essential component of urban geo-spatial data construction. Geographical name matching aims to compare paired strings representing the same real-world location. Current geographical name matching methods rely on either independent string similarity or a combination of multiple similarity metrics, which fail to effectively capture character substitutions involved in geographical name changes due to language and cultural variations. We proposed a geographical name matching method based on XLNet algorithm, which using a deep neural network to classify a pair of geographical name as match or non-match. The method based on long-term memory uses bidirectional information flow attention masks to reconstruct event sequences, establishing representations by using the bidirectional information of sequence. The experimental result demonstrates the effectiveness of this method in addressing the issue of lengthy address matching. The model can more comprehensively capture the semantic information conveyed within the context, which outperforms previous studies on single similarity metrics and supervised machine learning methods.

Keywords: geographical name matching, geographical name entity, XLNet, Softmax regression model

地名匹配是将代表同一现实世界位置的字符串进行配对，是与地理信息检索和地理信息科学相关的若干应用基本问题，如地理编码和地图搜索服务[1]或文本内容上的地名解析[2]，地名信息采集系统[3]和地名地址数据库[4]的建立。

早期的地名匹配方法从字符相似度的角度探究地名匹配问题，如函数法、全字匹配法、模糊查询法等的地名匹配算法[5]。此外，还有学者从空间几何的角度研究地名匹配的方法，以距离、位置等空间几何特性作为地址特征构造模型，但该方法涉及大量几何运算，执行效率低[6]。主流的地名匹配算法研究主要依赖深度学习模型。

目前基于深度学习的地名匹配算法通常采用序列到序列（Seq2Seq）模型[7]。然而单一的Seq2Seq模型可能无法满足所有的地名匹配需求。因此，研究人员不断探索改进模型和算法，例如使用注意力机制来更

收稿日期：2023-04-01。
项目来源：国家重点研发计划资助项目(2022YFB3904200、2022YFF0711601)；湖北省自然科学基金资助项目（2022CFB640）；地质探测与评估教育部重点实验室主任基金资助项目（GLAB2023ZR01）。
第一作者简介：郑诗语（1998—），硕士研究生，研究方向为地名匹配与时空知识图谱，E-mail：1202210925zsy@cug.edu.cn。
通信作者：邱芹军（1988—），博士研究生，副研究员，研究方向为地质大数据挖掘与信息抽取，E-mail：qiuqinjun@cug.edu.cn。

好地处理长序列，使用基于双向递归神经网络的混合神经网络来处理非结构句子，使用面向时空的卷积神经网络（SP-CNN）[8]或神经网络空间抽取模型（NeuroSPE），有效捕捉到自然语言空间关系的特征，使用多层编码器和解码器来处理更复杂的地名等。

目前的地名匹配技术仍存在着语义表征不充分的问题，鉴于此，本文提出一种基于XLNet算法的地名匹配框架。XLNet采用从左至右的AR方式输入数据，但在模型内部，通过注意力掩码方法来实现AE，从而充分利用了序列的双向信息。

1 概 述

地名匹配算法可分为：基于字符串相似度、基于空间几何角度、基于语义相似度三类。

基于字符串相似度匹配的方法可分为：基于字符、空间向量和混合方法。

基于空间几何角度的地名匹配算法关注匹配算法中的空间信息指标。SAALFELD将蜘蛛编码和点几何位置结合，提出点匹配方法[9]。Gabay[10]等提出利用点距与线段方向进行线串匹配的方法。陶云[11]提出一种几何匹配方法，通过对面实体进行栅格化和收缩，获得其质心并进行向量化。该方法易受空间数据精度、存取效率等因素的影响，普适性不高且运行效率较低，因此在实际地名匹配中不常用。

基于语义相似度的算法主要有基于语料库和基于知识库方法。基于语料库的地名匹配将地址与语料库中的信息相连接，从而增强地址信息量，获得更多语义特征，包括基于词袋模型和基于深度学习的方法。如Salton提出的空间向量模型[12]。该方法将文本处理转化为对向量的运算。基于深度网络的匹配方法主要是通过将地址文本输入神经网络模型中生成词向量，再计算文本相似度。

基于知识库的地名匹配利用知识图谱技术进行地名匹配。为获取地址的隐藏语义，常用的方法是将知识图谱嵌入实体中。如陈雨晖[13]等提出的一种基于知识图谱的中文地址匹配法，以三元组的形式存储地址元素及其关系，通过构建匹配地址知识图谱实现对地址信息的匹配。

2 基于XLNet算法的地名匹配框架

本文将XLNet算法与地名实体匹配相关理论结合，提出一种新的地名匹配模型。该模型首先对输入的中文地址对进行数据清洗，接着将处理好的待匹配数据输入双流自注意力模型进行特征抽取生成地址向量，后使用准备好的训练数据对XLNet模型进行训练，使用Softmax函数和损失函数来优化模型参数，使其能够准确地预测文本匹配情况。

2.1 数据预处理

本文提取高德地图的POI数据作为原始数据，该POI数据较为规整，因此只需进行繁简转换，爬取的POI数据以csv格式存入文件中。后基于处理过的地址集创建带标签的地址数据集作为训练集和测试集，生成带标签数据集的步骤如下：

1）标记不匹配的数据集：随机选取地址数据集中的一条数据，作为待匹配地址，标记为0（不匹配）。

2）标记匹配的数据集：选取一条地址记录作为目的地址，模拟人们在真实场景中检索的非标准模糊的地址，将目的地址转化为待匹配地址，标记为1（匹配）。转化的具体方法有如下几种：

删除元素：删除地址省，市，区，门牌号等信息。

添加元素：添加冗余信息，如“...附近”

最终数据库中匹配：不匹配地址数量比为1∶1。

2.2 XLNet模型

经过预处理的地址数据集需进行向量化，才能输入到XLNet匹配模型中进行匹配。传统的向量化方法使用Word2vec模型，但无法处理多义词问题，会造成地名匹配过程中无法区别同名不同地现象。因此本文使用XLNet模型进行文本向量表示。

XLNet模型基于AR(自回归)模型和AE(自编码）模型，其保留了AR模型的优点，并结合BERT模型（一种降噪自编码模型）原理，弥补自回归模型只考虑单方向语义的缺点，使模型能兼顾上下文信息[14]。

2.2.1 排列语言模型

XLNet模型借鉴神经自回归分布估计NADE[15]的思想，提出排列语言模型来弥补AR模型无法捕捉双向语境的缺点。排列语言模型的公式如式（1）：

$$\max_{\theta} E_{Z\sim Z_T}\left[\sum_{t=1}^{T}\log p\theta\left(X_{zt}\middle|X_{z<t}\right)\right] \tag{1}$$

式中，X 为文本序列；Z_T 为长度为 T 的索引序列[1，2，....，T]的所有可能排列的集合；z_t 为第 t 个元素的索引，$z<t$ 为一个序列的前 $t-1$ 个元素的索引。图1解释排列语言模型实现的过程：

输入：XLNet语言模型预训练仍然采用自回归模型的模式，因此句子依然按从左至右{ Z_1, Z_2, Z_3, Z_4 }的顺序作为输入，为了避免出现BERT模型中由于随机

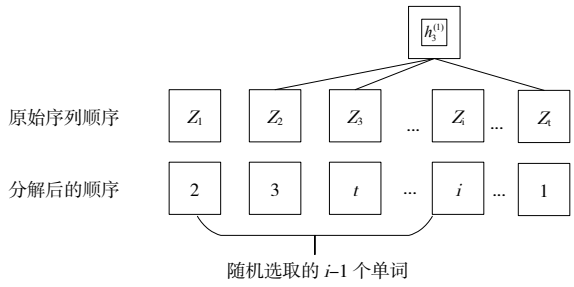


图 1　排列语言模型图

Mask 部分单词而导致预训练阶段和微调阶段不一致的问题，因此不能直接将排列组合作为原始输入。

注意力掩码：XLNet 采用 Transformer 中注意力掩码的方式，随机选取 $i-1$ 个单词作为被预测的第 i 个单词的上文，其余未被选中的单词被掩码隐藏掉，达到兼顾上下文进行预测的目的。

2.2.2　双流自注意力机制

在具体实现排列语言建模时，若采用传统的 Transformer 结构，则会导致预测时忽略上文语序的问题，这是由于传统的 Transformer 结构只考虑了词的内容，而没有考虑预测词的位置信息，这个缺陷可能导致模型得出不正确的表示结果。为解决这一问题，XLNet 采用双流自注意力模型，同时感知预测词的位置和内容信息。

双流自注意力模型中的“双流自注意力”指的是内容流自注意力和查询流注意力。内容信息流是对上下文信息的表示，包含当前词，查询信息流包含除当前此以外的上下文信息和当前词的位置信息，但不包含当前词的内容信息。在预测其他单词，内容信息流中可以看到其本身的信息，例如输入序列为$\{X_1, X_2, X_4, X_3\}$，随机生成的序列为$\{X_1, X_2, X_4, X_3\}$，XLNet 模型中加掩码形成的内容信息流矩阵如图 2 所示，如图对于排列后的 X_4，只能看到 X_1，X_2 的信息，因此在矩阵中表现为第 4 行的第 3 个位置的信息被隐藏掉。

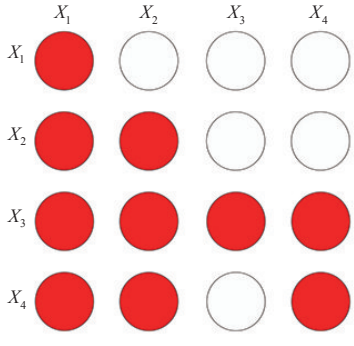


图 2　内容信息流矩阵图

在预测当前单词时，查询信息流中不能看到其本身的内容信息，因此 XLNet 模型中加掩码形成的查询信息流矩阵如图 3 所示，对于排列后的 X_1，位于第一个位置没有上文，因此在矩阵中表现为第 1 行的所有信息被隐藏掉。

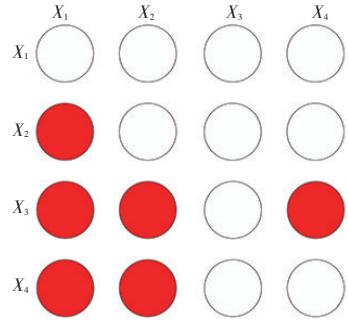


图 3　查询信息流矩阵图

结合上述 2 种信息流，XLNet 可以同时考虑预测词的位置信息和上下文信息，避免了传统 Transformer 结构的缺点。然而，序列的排列组合会增加数据量，导致计算量过大，使得模型开始训练时收敛缓慢。为了降低优化难度，本文选择一个超参数 K，只对最后的 $\frac{1}{K}$ 个词进行预测。

2.2.3　Transformer–XL

除了上述 2 个关键技术，XLNet 还采用最先进的 AR 模型 Transformer–XL 来解决长文本问题。

Transformer–XL 的相对位置编码机制是对 BERT 模型中绝对位置编码的改进。绝对位置编码若应用于分段情况则会导致位置信息的缺失，即会出现不同的片段相同词的向量编码表示相同的情况。针对该问题，相对位置编码机制提出了解决方法。对于序列中位置 i 和 j，若来自同一段，则标记为段编码 $S_{ij}=S_+$，若来自不同段，则为 $S_{ij}=S_-$，S_+ 和 S_- 为注意力机制中的可学习参数。相对位置编码可用公式表示为如式 2。

$$A_{i,j}^{\mathrm{rel}}=\underbrace{\boldsymbol{E}_{x_i}^{\mathrm{T}}\boldsymbol{W}_q^{\mathrm{T}}\boldsymbol{W}_{K,E}\boldsymbol{E}_{xj}}_{(1)}+\underbrace{\boldsymbol{E}_{x_i}^{\mathrm{T}}\boldsymbol{W}_q^{\mathrm{T}}\boldsymbol{W}_{K,R}R_{i-j}}_{(2)}+\underbrace{u^{\mathrm{T}}\boldsymbol{W}_{K,E}\boldsymbol{E}_{xj}}_{(3)}+\underbrace{v^{\mathrm{T}}\boldsymbol{W}_{K,R}R_{i-j}}_{(4)} \tag{2}$$

式中，$\boldsymbol{E}_{x_i}$，$\boldsymbol{E}_{xj}$为位置为 i，j 的词内容向量；$\boldsymbol{R}_{i-j}$ 为位置 i，j 的相对位置向量；$\boldsymbol{W}_q^{\mathrm{T}}$ 为权重变化矩阵；$\boldsymbol{W}_{k,E}$ 为学习到的基于内容的和基于位置的权重变换向量；$\boldsymbol{u}^{\mathrm{T}}$ 和 $\boldsymbol{v}^{\mathrm{T}}$ 为需要学习的参数向量。公式由 4 部分组成，①表示基于内容的“寻址”；②表示基于内容的位置偏移，即相对于当前预测词的位置偏差；③表示全局的内容偏移，衡量输入的上下文信息词的重要性；④表示全局的位置偏移，根据预测词和上下文信息的距离调整其权重。

2.3 Softmax回归函数

地名匹配任务的本质分类任务，即将2个地址的相似度分为匹配和不匹配两类，因此，在利用XLNet模型提取文本特征并得到文本向量表示后，需要将这些向量输入到Softmax回归函数中进行分类，从而得出2个地址间是否存在匹配关系的结果。

3 实验与分析

3.1 数 据

本文从高德地图中爬取了10 000条湖北省POI数据，其中2 000条武汉市，2 000条深圳市POI数据作为原始数据集。通过程序自动化制作带标签的地址数据集，将匹配的标签标注为1，不匹配的标签标注为0，最终得到30 000条湖北省POI数据对，6 000条武汉，6 000条深圳POI数据对。

20 000条湖北省POI数据对作为训练集训练所有模型。为了对本文提出的XLNet模型进行有效性分析，将6 000条武汉POI数据对作为测试集1输入XLNe模型以及其他对比模型，由此验证本模型相对于其他模型的优势。此外，为了验证XLNet算法在长地址匹配问题上的有效性，从6 000条深圳数据中选取2 000条地址长度大于平均值60%（实验中为字符串长度大于24）的数据作为长地址作为测试集2与其他模型进行对比实验。

3.2 实验环境

实验环境如表1所示。

表1 环境配置

名称	配置
CPU	11th Gen Intel(R) Core(TM) i7-11700 @ 2.50 GHz
硬盘	5.5 T
内存	64 GB
软件环境	pycharm (transformers 4.24.0 pytorch 1.8.2)
python	python 3.7

3.3 实验结果与分析

为评估基于XLNet算法的地名匹配模型性能，将所提出的方法与现有几种地名匹配方法进行比较，并使用准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1-score以及批处理时间作为评价指标衡量模型的性能。表2展示了各方法在测试集1获得的地名匹配精度结果。由于本文所制作数据集方法仅仅从地址字符增减角度进行制作，在提高数据集制作速度，降低人工工作量的同时也降低了数据集的多样性，因此模型所得结果精度普遍很高，为了更直观展现模型优劣，下面对数据进行详细的横向对比分析。

表2 不同地名匹配方法的比较评价

模型	Accuracy	Recall	Precision	*F*1	批处理时间/s
ABCNN	0.923 6	0.935 4	0.986 6	0.960 3	0.092 6
Albert	0.999 5	0.999 5	1	0.999 7	0.035 5
Bert	0.999 5	0.999 7	0.999 8	0.999 7	0.034 8
BIMPM	0.999	0.999	1	0.999 5	0.052 9
BIMPM	0.999	0.999	1	0.999 5	0.052 9
Decomposable Attention	0.990 6	0.990 7	1	0.995 3	0.015 9
SiaGRU	0.847	0.883 8	0.953 1	0.917 2	0.010 5
Robert	0.999 2	0.999 2	1	0.999 6	0.034 8
XLNet	0.999 6	0.999 7	1	0.999 8	0.056 4
Distillbert	0.838 3	0.838 3	1	0.912 1	0.019 2
ESIM	0.959	0.959	1	0.979 1	0.013

在原理上，这些模型可以分为基于自注意力和非自注意力机制两大类。其中，Albert、Bert、XLNet、DistilBert、Robert为自注意力模型。通过对比两类模型在不同评估指标下的表现，可以看出自注意力模型在地名匹配问题上的Accuracy，Precision，Recall，*F*1值均表现较好。自注意力模型之所以能够在建模序列数据时表现出色，是因为它能够关注输入序列中不同位置之间的相互关系，更好地捕捉序列中的长程依赖关系。因此，本文认为在地名匹配问题中，使用XLNet这种自注意力模型是一种较为可靠的选择。

图4中展示了4个模型在Accuracy，Precision，Recall，*F*1方面的性能对比，其中包括3个基于自注意力机制的模型和本文提出的XLNet模型。从图中可以明显看出，XLNet在所有指标上的表现都优于其他3个自注意力模型。这一结果表明XLNet模型在地面匹配问题上具有出色的性能。

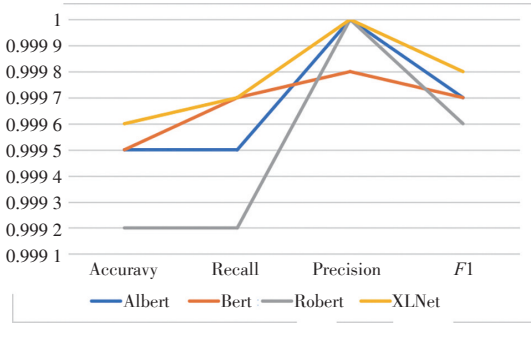


图4 自注意力模型性能对比

然而，尽管XLNet地名匹配算法在性能表现上优异，但其处理时间相对较长。如图5所示，各算法的批处理时间存在一定的差异，其中SiaGRU的批处理

时间最短，而XLNet算法的批处理时间较长。分析原因是XLNet算法采用了排列语言模型，增加了需要处理的数据量，导致批处理时间高于平均水平。相比之下，SiaGRU模型采用序列级别的计算方式，避免了序列对齐中矩阵级别的计算，因此具有较高的计算效率。后续可考虑采用并行计算的技术来加速模型的训练和推理过程，如多GPU训练。

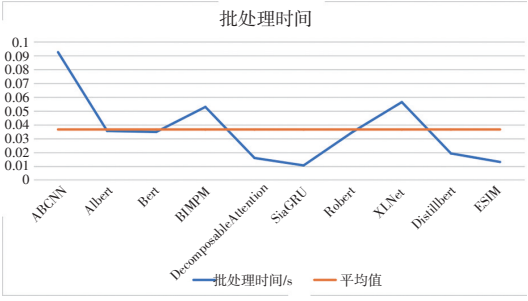


图5　不同地名匹配方法的批处理时间折线图

表3列举了XLNet模型在地名匹配任务上的测试结果，可以看出在长文本地址问题上，XLNet模型表现较为优秀。为了验证该结论，本文使用针对长地址对匹配使用2 000条深圳市长地址对进行对比实验，实验结果如表4，验证了上述结论。

表3　XLNet测试结果示例表

地址对	正确结果	输出结果
武汉市江汉区长港路325号/长港路325号	1	1
武汉市江岸区南京路107号（循礼门地铁站h口步行200 m）/武汉市青山区临江大道与三弓路交叉路口往西南约250	0	0
武汉市武昌区4号线；7号线/武汉市武昌区	1	0
武汉市洪山区珞喻路630号/珞喻路856号	0	1

表4　长地址匹配结果

模型	Accuracy	Recall	Precision	*F*1	批处理时间/s
ABCNN	0.671 5	0.699 5	0.943 8	0.803 5	0.039 5
Albert	0.988 5	0.998 7	0.998	0.994 2	0.036 4
Bert	0.987	0.993 5	0.993 5	0.993 5	0.035 6
BIMPM	0.975 5	0.977	0.998 5	0.987 6	0.047 8
Decomposable Attention	0.513 5	0.513 8	0.999	0.678 6	0.032 0
SiaGRU	0.5	0.5	1	0.666 7	0.010 6
Robert	0.989 5	0.996	0.993 5	0.994 7	0.035 7
XLNet	0.990 5	0.998	0.992 5	0.995 2	0.055 9
Distillbert	0.5	0.5	1	0.666 7	0.021 4
ESIM	0.902	0.902	1	0.948 5	0.168 2

但表3匹配错误数据结果同时可以看出，模型对于短地址数据处理结果不佳。这是由于上下文信息较少，模型容易出现过拟合或欠拟合的情况，从而导致分类效果下降，这也是XLNet模型在处理短文本数据时存在的一些缺点和挑战。

4　结　语

本文提出一种基于XLNet的新的地名匹配算法，该算法利用XLNet模型进行地名实体识别，捕捉地名信息特征，从而计算地名的匹配程度得到匹配结果。通过对比实验和评估结果表明，该算法较其他算法对上下文的理解更全面，能获得更准确的语义表征，在地名匹配任务上有较高的准确率和鲁棒性，并且能更好地处理长地址匹配任务。但本文的研究也存在一定的局限性，未来的研究可以考虑采用分布式训练、量化模型等方法提高模型数据处理效率。

参考文献

[1] Berkhin P, Evans M R, Teodorescu F, et al. A New Approach to Geocoding:Binggc[C]//Proceedings of the 23rd Sigspatial International Conference on Advances in Geographic Information Systems,2015:1-10

[2] Monteiro B R,Davis J C A,Fonseca F. A Survey on the Geographic Scope of Textual Documents[J]. Computers & Geosciences,2016,96(12):23-34

[3] 陶彧君,熊忠招,龚元夫,等.基于GIS技术的地名信息采集系统设计[J].地理空间信息,2020,18(11):74-76

[4] 雷双友,王志成,龚玉叶,等.标准地名地址数据库建设及应用[J].地理空间信息,2021,19(6):132-134

[5] 程钢,卢小平.顾及通名语义的汉语地名相似度匹配算法[J].测绘学报,2014,43(4):404-410

[6] 郝燕玲,唐文静,赵玉新,等.基于空间相似性的面实体匹配算法研究[J].测绘学报,2008,37(4):501-506

[7] Qiu Q,Xie Z,Ma K,et al. Spatially Oriented Convolutional Neural Network for Spatial Relation Extraction from Natural Language Texts[J]. Transactions in GIS, 2022, 26(2): 839-866

[8] Santos R, Murrieta-flores P, Calado P, et al. Toponym Matching Through Deep Neural Networks[J]. International Journal of Geographical Information Science, 2018, 32(2): 324-348

[9] Gabay Y, Doytsher Y. Automatic Adjustment of Line Maps[J].Proceedings of GIS/LIS,1994(94):332-340

[10] Yuan S,Tao C. Development of Conflation Components[J]. Proceedings of Geoinformatics,1999(99):1-13

（下转第88页）