arXiv:2301.01170v1 [cs.CL] 2 Jan 2023

基于Transformer的地理编码

Yuval Solaz* Vitaly Shalumov*
yuval.solaz@gmail.com vitaly.shalumov@gmail.com

2023年1月4日

摘要

本文将根据自由文本预测地理位置的问题表述为序列到序列问题。基于这一表述，我们通过训练一个T5编码器-解码器Transformer模型来获得地理编码模型，该模型以自由文本为输入，以地理位置为输出。该地理编码模型使用带地理标签的维基转储数据进行训练，并采用自适应网格划分来表示地理位置。本研究使用的所有代码，包括基于REST的应用程序、数据集和模型检查点，均已公开。

1 引言

Twitter 和 Wikipedia 等社交媒体包含大量与位置相关的文本数据。本文提出了一种模型，能够从自由文本中学习预测空间概率。给定一个查询句子，该模型会在地球表面上输出一个离散概率分布，为每个地理网格单元分配一个可能性，表示输入文本与该单元所涵盖位置相关的可能性。由此得到的模型能够定位种类繁多的句子。将此任务视为层次分类问题，使模型能够表达其对文本所关联位置的不确定性。该模型可用于消解文本中位置指称的歧义。这一能力对于从自由文本中找到精确位置至关重要。例如，巴黎 可能指多个不同地点。在如下语境中：国际奥林匹克委员会确认了*2024*年奥运会的举办城市。奥运会将在巴黎举行，本文提出的这类地理编码模型有助于确定正确的位置。

本研究的贡献如下：

- 合成用于监督学习的数据集，其中包括自适应网格划分。
- 将地理编码问题表述为序列到序列问题。
- 使用上述表述训练端到端地理编码模型。
- 公开发布整理后的数据集、基于REST的应用程序和 T5 地理编码模型。

*贡献相同

2 相关工作

文献中已广泛研究如何根据自由文本预测地理位置。 [1]的作者提出了一种多层地理编码模型，可学习将文本与地理位置关联起来。下游任务被表述为一个多层分类问题，其输出空间由多头模型中的多层 S2 [2] 单元格构成。在 [3], 中，地球表面被划分为数千个多尺度地理单元格。作者使用数百万张带地理标记的图像训练了一个深度神经网络。我们采用了本文提出的自适应划分方法。 [4] 的作者利用从带地理标记的 Twitter 数据中提取的坐标，构建了位置语言模型。这些位置按不同粒度建模，范围从邮政编码到国家层级。 [5], 提出了一种用于文本地理编码的端到端概率模型。此外，作者还收集了一个全新的数据集，用于评估地理编码系统的性能。他们将名为 ELECTRo-map 的基于模型的解决方案，与论文发表时可用的开源系统进行了比较，以对事件数据中的文本进行地理编码。 [6]提供了一种算法，可通过纯数据驱动的场景匹配方法，根据单张图像估算地理位置的分布。

上述地理编码方法可以视为由词元分类（假设该任务能够完美解决）和地理编码组成的序列任务。遗憾的是，这种方法存在若干缺陷。首先，它无法正确处理包含多个地点描述的输入（例如：“我们住在 X 国 Y 市”）。其次，它缺乏对自由文本上下文的理解（例如：“法国南方的国家”）。

3 基于 Transformer 的文本地理定位

我们将文本地理定位任务表述为一个序列到序列问题。模型将输入文本转换为一组具有层级结构的地理单元格，这些单元格表示地球表面上的概率分布。模型的输出是对地球表面层级单元格表示进行编码的序列。

3.1 自适应网格划分

我们使用谷歌开源的 S2 几何库 [2] 将地球表面划分为互不重叠的单元格，以此定义模型的目标。S2 几何库通过将外接立方体的各个面投影到球体上，对球体表面进行分层划分。立方体的六个面由六棵四叉树进行层级细分。四叉树中的一个节点定义了球体上的一个区域，称为 S2 单元。图 1 展示了不同分辨率下的 S2 单元。表 1 列出了 S2 几何各层级对应的分辨率和单元格数量。

选择这种细分方案而非简单细分纬度/经度坐标，有几个原因。首先，纬度/经度区域在两极附近会变得狭长，而 S2 单元则保持接近正方形的形状；其次，S2 单元的大小大多较为均匀（最大与最小 S2 单元的比例为 2.08）。定义地球平铺的一种简单方法，是使用层级结构中某个固定深度的所有 S2 单元，从而得到一组大小大致相同的单元格。然而，由于维基数据条目的地理分布 [7]（本文采用的数据集）在人口稠密地区呈现明显峰值，这种方法会导致类别分布极不平衡，见图 2。

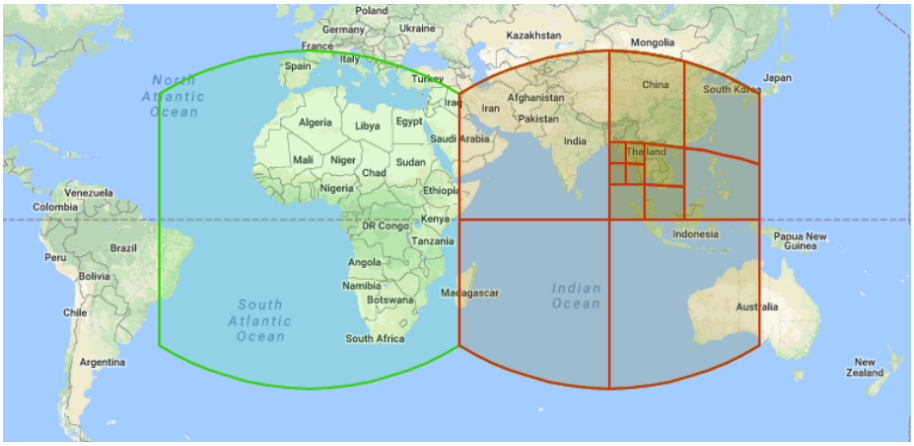

图 1：S2 几何对地球的分层划分。

表 1：S2 几何层级。

层级	平均面积	单元格数量
00	85M km^2	6
01	2100万 km^2	24
02	500万 km^2	96
03	130万 km^2	384
04	33万 km^2	1536
05	8.3万 km^2	6K
06	2万 km^2	24K
07	5K km^2	98K
08	1297 km^2	393K
09	324 km^2	1573
10	81 km^2	6M
..	..	..
29	2.95 cm^2	$1729 * 10^{15}$
30	0.74 cm^2	$7 * 10^{18}$

因此，我们根据数据集条目位置执行自适应细分：从根节点开始，递归遍历每棵四叉树，并细分单元格，直到没有单元格包含超过某个固定数量（单元格样本数上限参数）的数据点。采用这种方法后，人口稀疏地区由较大的单元格覆盖，而人口稠密地区则由更细的单元格覆盖。与均匀平铺相比，这种自适应平铺有几个优点：（i）训练类别更加均衡；（ii）能有效利用参数空间，因为会将更多模型容量用于人口稠密地区。图 3 展示了我们数据集的 S2 分区。

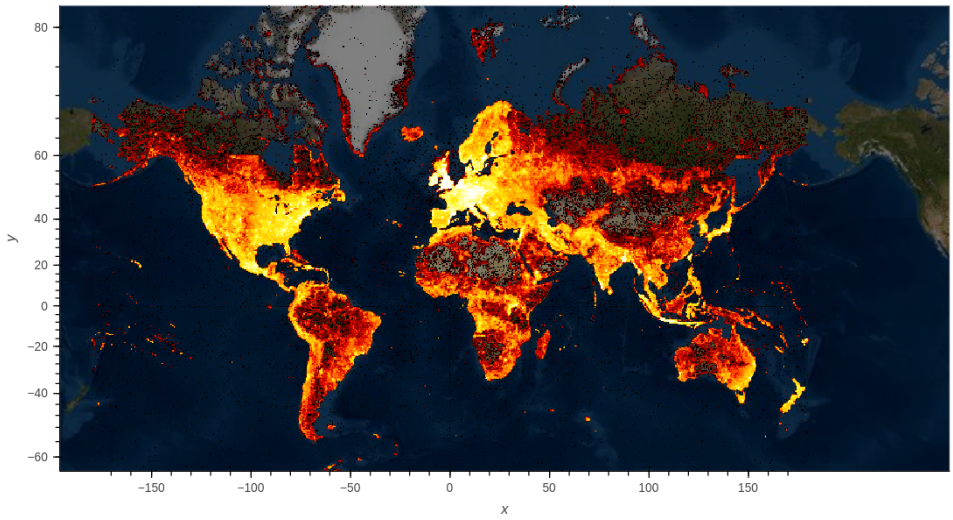

图 2：数据点分布。

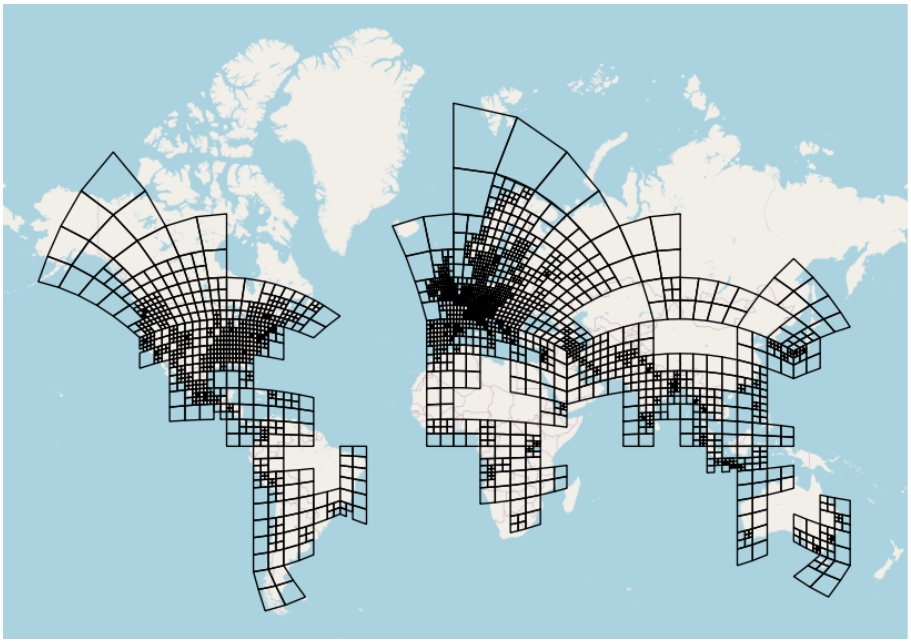

图 3：S2 几何对我们的数据集进行自适应分区。

3.2 数据集

本节介绍用于模型训练和评估实验的数据集。该数据集基于维基数据档案 [8] 构建。我们对档案进行筛选，选出所有带有位置标签的记录（约 800 万条记录）。参见表 2

适用于维基数据记录样本。

表 2：维基百科地理数据样本。

Id	纬度	经度	Text
01	53.96	-1.08	英格兰的历史郡
02	51.0	10.0	中欧国家
03	35.88	14.5	南欧的主权国家
04	57.30	-6.36	位于苏格兰高地、英国的威士忌酒厂
05	47.39	0.69	法国中央-卢瓦尔河谷大区安德尔-卢瓦尔省的一座城市和市镇
06	-33.0	-71.0	南美洲的主权国家

3.3 数据标注

使用自适应单元，我们为每个数据样本标注包含该样本位置的单元 ID。为了在标签中保留单元格的层级结构，我们采用以下单元 ID 编码：第一位表示单元立方体面，取值为*0* 至*5*之间的数字。后续各位分别表示四叉树中每一层对应的节点，取值为 *0* 至 *3*之间的数字。单元编码示例见表 3。请注意，标签是由数字组成的序列，长度不固定，介于*1* 和 *1+*最大层级之间。

表 3：单元编码。

单元描述	单元表示
面单元 2	2
面单元 1 的子单元 2	12
面 4 的子单元 3 的子单元 1	431

4 模型

4.1 训练

本文提出的模型基于 T5-base（220M）预训练序列到序列模型 [9]。我们使用文本记录作为输入、位置单元编码作为输出目标序列，在 Wikidata 数据集上对模型进行了微调。微调过程中，我们采用标准交叉熵损失。损失函数中基于层级的加权评估留待未来工作。我们训练了 5 个 epoch，共 2M 步，批量大小为 12。我们使用 AdamW 优化器，学习率为 1.5e-5，并采用线性衰减。我们使用 wiki-data 数据集的 80% 训练模型，其余 20% 用于域内评估。图 4 展示了文本到位置框架的示意图及若干输入/输出示例。

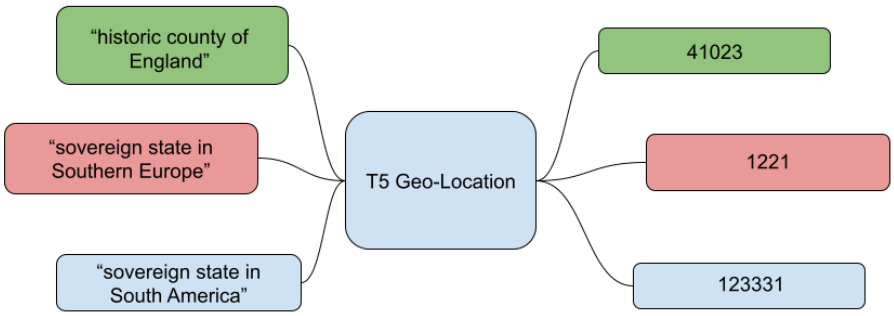

图 4：我们的文本到位置框架示意图。



图 5：文本“海法”的预测结果。蓝色为预测的 S2 单元，白色为其祖先单元格。

4.2 评估

4.2.1 推理

我们所得模型的推理过程如下：给定一个句子，我们预测输出序列。序列自然会以 1 个介于 0 到 5 之间的数字表示预测的立方体面，后面最多跟 9 个介于 0 到 3 之间的数字，分别表示各层级上预测的 S2 单元。该序列可以轻松转换为一个 S2 单元，它表示输入文本位置在地球上的概率分布。我们采用了大小为 10 的束搜索。推理示例见图 5。

4.2.2 评估指标

最直观的分类指标是“准确率度量”：只有当模型输出与真实标签完全匹配时，预测才算成功。我们将这一指标称为“平面准确率”（与“层次准确率”相对）。

这一指标未能体现标签固有的层次结构。另一种度量是“平均距离误差”。平均距离误差计算预测位置（在我们的例子中，即预测 S2 单元的中心）与目标文本真实位置之间的距离的平均值。它同样未能体现标签的层次结构。

因此，我们倾向于采用《常规分类指标的层次特定变体》 [10]中所述的层次分类指标。这些指标是广为人知的精确率、召回率和 F 分数指标的变体，专门针对层次分类进行了调整：

层级精确率 (hP)：

$$hP = \frac{\sum_i |P_i \cap T_i|}{\sum_i |P_i|}, \tag{1}$$

层级召回率 (hR)：

$$hR = \frac{\sum_i |P_i \cap T_i|}{\sum_i |T_i|}, \tag{2}$$

以及层级 F 度量 (hF)：

$$hF = \frac{2 * hP * hR}{hP + hR} \tag{3}$$

其中， P_i 是由每个测试样例 i预测出的最具体类别及其所有祖先类别组成的集合。 T_i 是由测试样例i的真实最具体类别及其所有祖先类别组成的集合。每个求和均针对测试集中的所有样例计算。

5 结果

本节展示并分析所开发的地理编码模型的性能。为展示模型的性能，我们先来看几个模型的推理示例。表 4 给出了预测样本的示例。

表 4：推理示例——真实标签和预测标签。

Text	预测标签	真实标签
爱尔兰克莱尔郡 Drummaan 的乡镇	21002321	21002321
瑞典 Eksj¨o 市镇的湖泊	20302303	20302303
丹麦的古代纪念碑（2976）	20331122	20331122
柴郡西和切斯特的学校 , UK	210033112	210033113
伊朗的山	1333313	133302
德国哈尔堡的铁路站	20331203	20331022

由于两个相邻单元格可能来自不同的父单元格，因此它们的序列在字符串比较方面可能差异很大，难以进行比较。为了直观呈现结果，我们构建了一个由微调模型驱动的网页演示应用程序。该演示应用程序有助于直观地比较预测结果与真实标签。图 6、图 7 和图 8 展示了使用该应用程序进行推理的示例。



图 6：文本“太平洋”的预测结果。

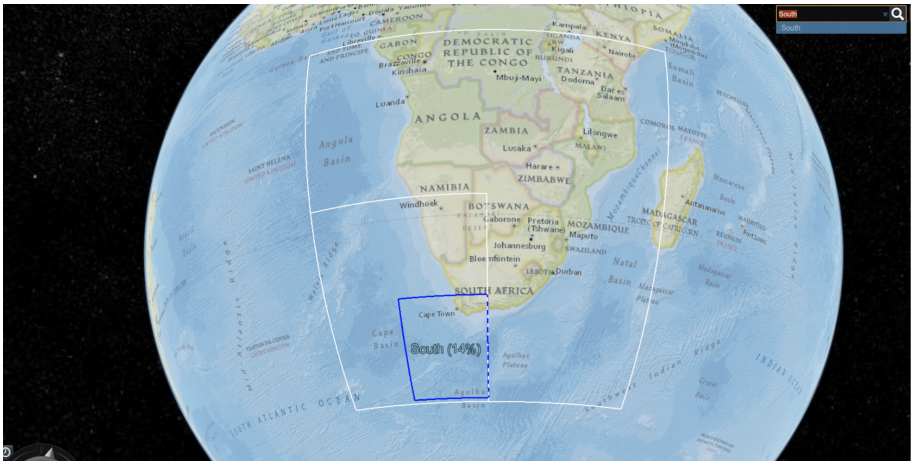

图 7：文本“南方”的预测结果。我们可以假设，大量南非样本使预测结果产生了偏差。

使用第 4.2.2 节所述指标，在维基数据测试集划分上得到的评估结果见表 5。

表 5：评估结果。

评估指标	结果
平面准确率	0.51547
层级准确率	0.791



图 8：文本“巴黎”的预测结果。

6 结论

本文将地理编码问题表述为序列到序列问题，并训练了一个用于地理编码的 Transformer 模型。为了充分利用语言模型的能力，我们使用预训练的 T5 模型进行微调。对于地理位置标签的序列表示，我们采用了自适应网格划分。自由文本及其对应的地理位置取自维基数据。模型在层次化指标和非层次化指标上的评估结果验证了所提出的地理位置预测方法的有效性。

推理时的自由文本有时只能间接暗示某个位置。我们认为，将 GPT3 [11]这样的大型解码器与 T5 结合，先根据模糊的位置指代生成更符合地理位置的文本，再将其输入 T5，或许能够提升模型性能。这项评估留待未来工作。此外，我们也将来再评估该方法在 Wikipedia 地名检索 [12]等基准上的表现。

参考文献

[1] Sayali Kulkarni、Shailee Jain、Mohammad Javad Hosseini、Jason Baldridge、Eugene Ie 和 Li Zhang。多层级地理编码的空间语言表示。*arXiv* 预印本 *arXiv:2008.09236*，2020年。[2]Dewey Dunnington、Edzer Pebesma 和 Ege Rubak。*s2*：使用 *S2* 几何库的球面几何运算符，2022年。https://r-spatial.github.io/s2/，https://github.com/r-spatial/s2，https://s2geometry.io/。[3] Tobias Weyand、Ilya Kostrikov 和 James Philbin。使用卷积神经网络进行行星照片地理定位。载于欧洲计算机视觉会议，第37–55页。施普林格，2016年。

[4] Sheila Kinsella、Vanessa Murdock 和 Neil O'Hare。"我正在格拉斯哥吃三明治"：利用推文对地点进行建模。载于第三届用户生成内容搜索与挖掘国际研讨会论文集，第61–68页，2011年。[5] Benjamin J Radford。基于文本回归地点以实现概率地理编码。*arXiv* 预印本 *arXiv:2107.00080*，2021年。[6] James Hays 和 Alexei A Efros。I m2gps：从单张图像估计地理信息。载于*2008*年 *IEEE*计算机视觉与模式识别会议，第1–8页。IEEE，2008年。[7] Denny Vrandeˇci´c 和 Markus Kr¨otzsch。维基数据：一个免费的协作知识库。*ACM*通讯，57(10)：78–85，2014年。[8] Denny Vrandeˇci´c 和 Markus Kr¨otzsch。维基数据：一个免费的协作知识库。*ACM*通讯，57(10)：78–85，2014年。[9] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li、Peter J Liu 等。探索统一文本到文本 Transformer 的迁移学习极限。*J. Mach. Learn. Res.*，21(140)：1–67，2020年。[10] Carlos Nascimento Silla 和 Alex Alves Freitas。跨不同应用领域的层次分类综述。数据挖掘与知识发现，22：31–72，2010年。[11] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少样本学习者。神经信息处理系统进展，33：1877–1901，2020年。[12] Milan Gritta、Mohammad Taher Pilehvar、Nut Limsopatham 和 Nigel Collier。地理解析中缺少什么？语言资源与评估，52(2)：603–623，2018年。