性别偏见研究中的代表性不足：基于异构图注意力网络的中文姓名性别预测

Zihao Pan[*]、**Kai Peng**[*]、**Shuai Ling**、**Haipeng Zhang** [†]

上海科技大学{panzh、pengkai、lingshuai、zhanghp}@shanghaitech.edu.cn

arXiv:2302.00419v1 [cs.AI] 1 Feb 2023

摘要

实现性别平等是人类可持续未来的重要支柱。开创性的数据驱动性别偏见研究基于大规模公共记录，例如科学论文、专利和公司注册记录，涵盖女性研究人员、发明者和企业家等群体。由于相关数据集中通常缺少性别信息，研究往往依赖工具根据姓名推断性别。然而，现有的开源中文性别推测工具尚不适用于科学研究，这可能是中文女性在主流性别偏见研究中代表性不足并影响研究普适性的部分原因。具体而言，这些工具侧重于字符级信息，却忽略了多字姓名中汉字的组合，以及汉字的部件和读音所传达的重要信息。作为首次尝试，我们设计了中文异构图注意力（CHGAT）模型，以捕捉部件关系中的异质性，并纳入汉字的读音。我们的模型大幅超越了现有工具，性能也优于当前最先进的算法。最后，最常用的中文姓名性别数据集基于单字，女性覆盖率低得多，且来源不可靠，这自然阻碍了相关研究。我们从官方来源获取了一个更加均衡的多字姓名数据集，并将其与代码一同开源，希望帮助未来的研究推动性别平等。

引言

近年来，基于科学论文、专利和公司注册记录等大规模公开记录，针对女性研究人员（Larivi`ere 等，2013；Huang 等，2020）、发明者（Jensen、Kov´acs 和 Sorenson，2018；Koning、Samila 和 Ferguson，2021）、企业家（Ritter-Hayashi、Vermeulen 和 Knoben，2019；van den Oever，2021）以及 STEM 学生（Cimpian、Kim 和 McDermott，2020）的性别平等研究日益增多。鉴于许多此类数据集不包含性别信息，研究者通常根据个人姓名推断性别。令人惊讶的是，中文女性在此类研究中的代表性不足——尽管中文领域存在大量可比数据。一个可能的原因是，与英文姓名性别推断工具相比，缺乏符合科学研究标准的可靠中文姓名性别推断工具。

Ngender[1] 是一种基于朴素贝叶斯的中文姓名性别预测基础工具。它通过将各个汉字的概率相乘，计算一个名字（通常由一到两个汉字组成）为女性名字的概率。尽管方法简单，Ngender仍有两项局限。其一与朴素贝叶斯本身有关——对于分类器从未见过的样本外字符，该方法无法处理。此外，它忽略了汉字组合以及汉字部件所蕴含的信息。Ngender训练数据也进一步加剧了这两项局限：数据只包含每个可用字符出现在女性姓名和男性姓名中的次数，而非完整名字在两种性别下的频率。已有研究证明，BERT和GloVe等神经网络语言模型生成的词表示具有性别倾向并传递性别信息（Jia 和 Zhao 2019年；Yang和Feng 2020；Lauscher等人 2020；Matthews、Hudzina和Sepehr 2022）。其中，Jia 和 Zhao（2019）提出了一种基于BERT的模型，该模型利用预训练BERT为每个汉字（无论是否在样本中出现）生成表示，从而解决姓名性别预测中的“字符样本外”问题。此外，通过拼接字符嵌入与发音嵌入，他们的模型拼音BERT（PBERT）也证明了读音能够传递性别信息。

然而，当前的性别推断工具忽略了由汉字部件产生的语义，而这些语义不仅来自汉字本身及其读音。大量汉字由部件构成，而这些部件有助于塑造汉字的含义（Yu等人，2017）。受此启发，研究者以word2vec的方式利用部件级内部语义特征进行汉字表示学习（Sun等人，2014；Yu等人，2017；Zhang等人，2019）。沿着这一思路，Wang 等人（2021）更进一步，通过构建同质图来捕捉部件相同的汉

[*]这些作者的贡献相同。

[†]*通讯作者。版权所有 © 2023，美国人工智能促进协会（www.aaai.org）。保留所有权利。

[1]https://github.com/observerss/ngender

字之间的语义关系。这样，汉字与其部件相互连接，借助注意力机制，更好地刻画并赋予汉字之间的语义关系以权重。尽管他们的FGAT模型在许多下游自然语言处理任务中达到了最先进水平，但其所依赖的关系实际上往往是异质的，而且尚未考虑同音连接；正如PBERT所示，这类连接有望增强图结构。

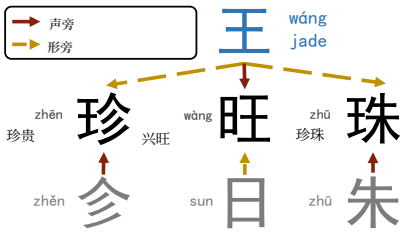

图1：共享部件连接示例。“ ”是“ ”和“ ”的形旁，但却是“ ”的声旁。

我们通过一个示例说明共享部件连接中的异质性，并讨论它如何影响我们的模型设计。在图1中，我们看到字符‘’（珍贵，*zhēn*）[2]，‘’（珍珠，*zhū*）和‘’（兴旺，*wàng*）共享部件王（玉，*wáng*）。如果像FGAT那样用同构图建模，就会认为‘’、‘’和‘’两两之间具有相同的语义相似度，这种相似度源自共享部件‘’。实际上，‘’和‘’在语义上要比它们与‘’更接近。就性别而言，‘’和‘’常见于女性名字，而‘’则有很强的男性倾向。因此，同质地建模它们之间的关系会误导姓名性别预测。事实上，如果我们知道‘’中的‘’是声旁（图1中的实线箭头），表示“旺”的发音，而‘’和‘’中的‘’是形旁（图1中的虚线箭头），有助于表达其含义，那么就能区分它们之间的关系。除了‘’，如图1底部所示，‘’还包含形旁‘’（太阳）。类似地，‘’和‘’分别以‘’（*zhū*，珠）和‘’（*zhěn*，珍）为声旁，兼有形旁和声旁的字符，例如‘’、‘’和‘’，称为形声字。80.5%的中文字符属于形声字（Sun 1997），这表明共享部件关系中的异质性影响不容忽视。因此，我们设计了异构图，以分别表示字符语义成分边和字符声旁边。

在这一模型结构中，鉴于读音在性别猜测中的有效性，我们引入共享读音连接，作为一种新型边。举一个简单的例子，共享“mei”这一读音的汉字（其读音与汉字“”相同，意为美丽）更有可能出现在女性名字中。现在，借助汉字-义符关系边、汉字-声旁边和汉字-读音边，我们首先使用多层级汉字注意力网络，在部件层面学习不同部件及结构信息的重要性，然后聚合读音（即拼音）所传达的性别信息。

除了方法上的贡献外，我们还提供了一个高质量的中文姓名性别数据集。大多数中文姓名性别数据集（例如Ngender 数据集）仅包含单个汉字的出现频率，而非完整姓名，因此会丢失中文姓名性别预测所需的重要信息。与通常由一个词构成的英文名字不同，大多数中文名字由一到两个汉字组成，其中双字名字占多数（84.55%）[3]。对于包含多个汉字的中文名字，汉字组合可能包含有用信息，有时甚至会传达与单个汉字所体现的信息相反的含义。例如，“”（胜）和“”（男）是更常出现在男性姓名中的汉字，但“”（胜过男性）却是女性名字。此外，正如我们在数据集章节中的胜楠所示，汉字组合 *A-B* 和 *B-A* 有时具有不同的性别概率。为了发挥序列表示的优势并推动相关研究，我们将从官方来源收集的大规模完整名字数据公开发布。

总而言之，本文的贡献包括：

1. 据我们所知，这是首项利用图神经网络对汉字的内外部联系进行建模，以实现姓名性别预测并促进性别平等研究的工作。
2. 我们提出了一种结合多级注意力网络的异构图，用于捕捉汉字与部件之间语义关系的异质性，以及读音所体现的性别倾向。该模型优于多种基线方法，达到了 93.62% 的最先进准确率。
3. 我们提供了一个包含 5800 万个带有对应性别信息的中文姓名的数据集，以及我们的源代码[4]，，希望借此推动性别平等研究，尤其是针对代表性不足的中国女性的研究。

相关工作

姓名性别预测

姓名性别预测是性别平等研究中的常见任务（Larivi`ere 等，2013；Huang 等，2020；Koning、Samila 和 Ferguson，2021），也是广告和推荐系统等网络服务中的常见任务（Mukherjee 和 Bala，2017；Wu 等，2019）。姓名在许多语言和文化中都包含丰富信息，是性别猜测的重要线索之一。

对于西方姓名，Wais（2016）设计了 genderizeR，仅根据数据中以某个输入名称命名的人群的

[2]含义和发音（斜体）位于括号内。

[3]www.mps.gov.cn/n2253534/n2253535/c8349222/content.html [4] https://github.com/ZhangDataLab/CHGAT

性别多数来预测该输入名称的性别。神经网络语言模型进一步有助于利用更多信息并解决“样本外”问题。Hu 等人（2021）构建了一个字符级基于BERT的模型，用于根据英文姓名推测性别。

中文不同于拉丁语系语言，是一种表意文字。除了汉字本身，汉字的部件和读音也能传递信息（Cao 等人，2018）。此外，在多字中文姓名中，汉字之间的组合还能带来额外信息。然而，大多数性别猜测工具只依赖字符级信息。例如，Ngender 和 Zhao 和 Kamareddine（2017）提出的模型，都将某一性别下所有汉字概率的乘积视为姓名属于该性别的概率，并输出概率最高的性别。这些方法在捕捉汉字组合所带来的额外信息方面存在固有局限。Jia 和 Zhao（2019）将预训练BERT模型中的字符嵌入和读音嵌入拼接起来，以缓解这些局限。然而，汉字间由共享部件和读音所体现的关联仍未得到充分利用。

用部件表示字符

汉字的部件蕴含丰富的语义信息。以往的研究基于 word2vec 模型，将部件信息融入字符嵌入，以学习汉字的表示（Sun等人，2014；Yin 等人，2016；Yu等人，2017）。为获取更细粒度的信息，一些研究将部件拆分为按序排列的笔画（子部件），以增强表示（Cao 等人，2018；Zhang等人，2019）。尽管汉字被拆解为部件或子部件，word2vec 模型仍无法区分它们的重要性，而无关部分可能会给表示引入噪声。

因此，Wang 等人（2021）在带有注意力机制的同构图中对字符及其部件进行建模，以学习部件的重要性。同时，通过字符与其他具有相同部件的字符之间的连接，该模型塑造了字符的语义。因此，其FGAT模型在各种下游自然语言处理任务上取得了最先进水平的结果。然而，正如引言部分所述，他们的同构图无法建模字符-部件关系中的异构性，因为大多数中文字符都是形声字，字符-形旁关系和字符-声旁关系应当分别以异构图的形式进行指定。此外，还可以将共享发音关系整合到异构图中，从而有望为我们的预测任务带来更多性别信息。

方法

我们的模型整体结构如图2所示。我们将汉字及其读音作为输入，以学习姓名中字符内部和读音组合内部的信息。以汉字形式呈现的输入姓名会同时输入中文异构图注意力（CHGAT）层和 BERT 文本编码器层。两个层输出的嵌入相加后，再与由 BERT 文本编码器层生成的姓名读音嵌入拼接。随后，将该嵌入输入 Transformer 编码器模块，以学习姓名中的上下文信息。最后，使用由单层全连接网络构成的分类器预测性别。接下来的章节将详细介绍异构图的设计及其核心组件 CHGAT 层。

异构图结构

汉字构形。 中文源自古代象形文字书写系统。木、火等基本字符最先出现，并各自具有图形表示。表示更复杂概念的字符由这些基本字符组合而成（Tao 等人（2019））。

例如，将两个“木（*mˉu*）”并列组合，得到“林（*lˊin*）”。将“林”（*lˊin*）”与其下方“火（*huˇo*）”组合，就会变成“焚（*fˊen*）”。

尽管汉字的字形会随时间变化，但其组合方式（即结构）得以保留。根据 QXK[5] （由北京师范大学研究人员构建的汉字信息系统），现代汉语中共有 17 种结构，我们在表1中列出了这些结构。正如引言中所述，80.5% 的简体汉字属于形声字，QXK 提供了大多数汉字的形旁和声旁。

图结构。 本文定义了一种异构图，包含四种类型的节点，即汉字、语义部件、声旁和发音，以及三种类型的路径，包括汉字-义符关系、汉字-声旁和汉字-发音，以融入性别信息。

如果一个汉字是形声字，它会通过汉字-声旁边与表示其读音的一个部件相连。对于其他部件，则通过汉字-义符关系边与之相连。如果一个汉字不是形声字，它也会通过汉字-义符关系边与其部件相连。此外，还会在汉字与其发音之间添加一条汉字-读音边。

我们以字符“ ”为例，说明在异构图中如何表示并连接汉字、读音和不同类型的部件。在图3中，焦点字符“ ”与其声旁“ ”通过汉字-声旁边相连，而字符“ ”与其形旁“ ”则通过汉字-义符关系边相连。

先前研究（Wang 等人（2021））利用同质图来表示部件与字符之间的关系，其中只有一种节点类型表示所有部件和字符，并且只有一种边类型

[5] https://qxk.bnu.edu.cn/#/help

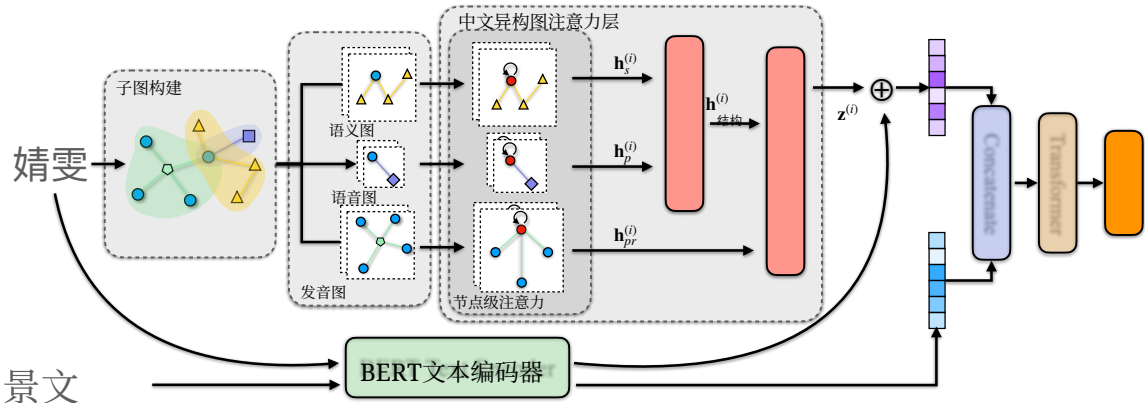

图2：系统结构。

索引	结构类型	Icon	示例
1	从左到右	⿰	桦, 婧
2	从左到 中间和右侧 ght	⿲	树, 辫
3	从上到下	⿱	雯, 芳
4	上、中、下	⿳	衷, 害
5	全包围	⿴	园, 围
6	从上方包围	⿵	阔, 闯
7	从下方包围	⿶	凶, 函
8	从左侧包围	⿷	区, 匠
9	从左上方包围	⿸	房, 庐
10	从右上方包围	⿹	勾, 氦
11	从左下方包围	⿺	赶, 迁
12	积分	□	一, 口
13	等腰三角形布局	△	鑫, 森
14	方框结构	⊕	品, 燚
15	多元素组合	◎	彝, 燮
16	叠置	§	爨, 南
17	多元素堆叠	※	爽, 坐

表1：中文汉字构形类型列表。前11种结构类型的名称沿用其 Unicode 名称，后6种没有正式名称，因此我们尽可能描述其结构特征。

用于字符间共享部件关系。他们假设同一字形作为字符、形旁或声旁时都表示相同的语义，这过于简化了。我们区分形旁和声旁，并将焦点字符与所有共享相同部件的字符相连。

如引言部分所述，Jia 和 Zhao（2019年）发现，汉字读音关系具有性别倾向，但他们仅将读音作为独立表示使用，没有将消息与汉字嵌入进行交互。因此，如图3所示，我们通过一条汉字-读音连接字符‘ ’与其发音‘*zh¯u*’。

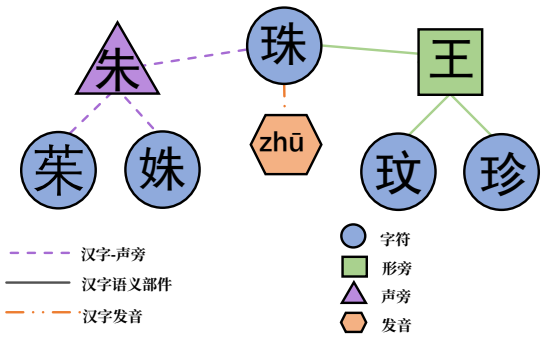

图3：字符“ ”的图构成示例。“ ”、“zh¯u”和“ ”分别是‘珠’的声旁、发音和语义部件，它们分别通过汉字-声旁、汉字读音关系和汉字语义部件与“ ”相连。

中文异构图注意力

为了聚合异构图中的信息，我们设计了具有三级注意力的中文异构图注意力（CHGAT）层。形状图和语音图都包含结构信息。节点级注意力聚合各图内部的信息后，结构信息仍分散在形旁和声旁节点中。因此，网络通过聚合两个节点级表示来捕获字符的结构信息。为了整合发音与结构表示中的信息，我们使用了另一种注意力机制，即聚合注意力。

我们将汉字集合记为 $\mathcal{C} = \{c_0, c_1, ..., c_m\}$。字符 c_i 的特征嵌入为 $\boldsymbol{f}_c^{(i)}$。所有汉字共享一个发音图 $\boldsymbol{g}_{pr}$，按照 Wang等人（2019年）的定义，它是一个“汉字-发音-汉字”元路径；但每个字符都有自己的语义图和语音图。对于字符 c_i，其语义图包含 c_i 的所有一跳和两跳语义部件，记为 $\boldsymbol{g}_s^{(i)}$；而其语音图

则包含 c_i 的一跳语音成分，表示为 $\boldsymbol{g}_p^{(i)}$。语义图中部件 s_o 的特征嵌入为 $\boldsymbol{f}_s^{(o)}$，而语音图中部件 p_t 的特征嵌入为 $\boldsymbol{f}_p^{(t)}$。

Wang 等人（2021）引入了位置嵌入，为成分表示添加位置信息。相同的部件可能出现在不同汉字的不同位置，而这些位置信息会影响部件的含义。因此，我们添加了另一种位置嵌入，用于表示每个部件在汉字中的位置，记作

$$\boldsymbol{\lambda} = \{\boldsymbol{\lambda}^{(0)}, \boldsymbol{\lambda}^{(1)}, ..., \boldsymbol{\lambda}^{(a)}\}. \tag{1}$$

因此，输入字符的初始嵌入 $\boldsymbol{x}_c^i$、语义部件的初始嵌入 $\boldsymbol{x}_s^o$ 和声旁的初始嵌入 $\boldsymbol{x}_p^t$ 分别为：

$$\boldsymbol{x}_c^i = \boldsymbol{f}_c^{(i)} + \boldsymbol{\lambda}^{(l_c)}, \tag{2}$$

$$\boldsymbol{x}_s^o = \boldsymbol{f}_s^{(o)} + \boldsymbol{\lambda}^{(l_s)}. \tag{3}$$

$$\boldsymbol{x}_p^t = \boldsymbol{f}_p^{(t)} + \boldsymbol{\lambda}^{(l_p)}, \tag{4}$$

分别如此。i、o 和 t 分别表示汉字、语义部件和声旁的索引。l_c、l_s 和 l_p 分别表示它们对应的位置索引。

在此，我们将汉字 c_i 的语音图特征嵌入定义为 $\boldsymbol{X}_p^{(i)}$，其语义图特征嵌入定义为 $\boldsymbol{X}_s^{(i)}$。发音图特征嵌入为 $\boldsymbol{X}_{pr}^{(i)}$。

受 HAN（Wang等人（2019））的启发，我们在此场景中采用三级注意力机制。如图2所示，模型首先学习每个图中的节点级嵌入，然后将语义图和语音图的节点级嵌入输入结构注意力，以学习结构嵌入。接着，利用该嵌入和节点级发音嵌入，通过聚合注意力学习汉字的最终嵌入。

节点级注意力。我们首先通过一个变换矩阵，将不同类型的节点特征投影到相同的特征空间中：

$$\boldsymbol{\gamma}_k^{(i)} = \boldsymbol{W}_k^{(i)} \boldsymbol{x}_k^{(i)}, \tag{5}$$

其中，$\boldsymbol{W}_k^{(i)} \in \mathbb{R}^{d_k \times d}$ 是一个可学习参数。$k \in \{s, p, pr\}$ 表示路径类型，d_k 是输入特征维度。节点 j 对目标节点 i的重要性分数计算如下：

$$n_k^{(i,j)} = \boldsymbol{LeakyReLU}(\boldsymbol{w}_k[\boldsymbol{\gamma}_k^{(i)} || \boldsymbol{\gamma}_k^{(j)}]), \tag{6}$$

其中，$\boldsymbol{w}_k \in \mathbb{R}^{1 \times 2d}$ 是一个可学习向量。随后对重要性分数进行归一化，得到节点 j 到节点 i 的权重，记为 $\theta_k^{(i,j)}$：

$$\theta_k^{(i,j)} = \frac{\exp(n_k^{(i,j)})}{\sum_{j \in \mathcal{N}_k^{(i)}} \exp(n_k^{(i,j)})}, \tag{7}$$

其中，$\mathcal{N}_k^{(i)}$ 是 k 类型路径中 i 的所有邻居集合。节点级嵌入则是所有与其相连节点的加权和：

$$\boldsymbol{h}_k^{(i)} = \Big\|_t \boldsymbol{ELU}\Big(\sum_{j \in \mathcal{N}_k^{(i)}} \theta^{(i,j)} \boldsymbol{\gamma}_k^{(i)}\Big), \tag{8}$$

其中，t 是注意力头的数量，$\|$ 表示拼接操作。

注意力模块。注意力模块将具有不同语义的表示聚合为一个表示。我们将其记为：

$$\boldsymbol{h}^{(i)} = attn(\boldsymbol{h}_1^{(i)}, \boldsymbol{h}_2^{(i)}, ..., \boldsymbol{h}_v^{(i)}), \tag{9}$$

其中，v 表示输入的数量。每个输入对目标嵌入的重要性为：

$$w_r = \frac{1}{|N|} \sum_{i \in N} \boldsymbol{q}^T (tanh(\boldsymbol{W}\boldsymbol{h}_r^{(i)} + \boldsymbol{b})), \tag{10}$$

其中，$\boldsymbol{q}$、$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 是模型中的可学习参数。$r \in \{1, 2, ..., v\}$ 是语义类型，N是输入名称。随后，每个输入的重要性分数为：

$$\delta r = \frac{exp(w_r)}{\sum_{u \in [1,v]} exp(w_u)}. \tag{11}$$

目标嵌入为：

$$\boldsymbol{h}^{(i)} = \sum_{r \in [1,v]} \delta r \boldsymbol{h}_r^{(i)}. \tag{12}$$

结构注意力层。在结构注意力层中，我们通过注意力模块，根据 c_i 在语义图和语音图中的节点级嵌入，学习其结构表示：

$$\boldsymbol{h}_{structure}^{(i)} = attn(\boldsymbol{h}_s^{(i)}, \boldsymbol{h}_p^{(i)}). \tag{13}$$

$\boldsymbol{h}_{structure}^{(i)}$ 表示字符 c_i的结构表示。

聚合注意力层。聚合注意力层将字符 c_i的发音表示与其结构表示整合为一个嵌入。同样，这是通过应用注意力模块实现的，具体如下：

$$\boldsymbol{z}^{(i)} = attn(\boldsymbol{h}_{structure}^{(i)}, \boldsymbol{h}_{pr}^{(i)}). \tag{14}$$

最后，$\boldsymbol{z}^{(i)}$ 是 c_i 在CHGAT层的输出。

损失函数

我们的目标函数定义如下：

$$L = -\sum_j^J y_j log(\hat{y}_j) + (1 - y_j) log(1 - \hat{y}_j) \tag{15}$$

其中，J 表示数据的数量，y_j 是j^{th} 输入的标签，而$\hat{}$ y_j 表示 y_j属于该标签的预测概率。

	记录数	不同名字	男性对女性比例
Ngender	32,067,566	9,442	197.28
我们 的数据集	58,393,173	560,706	111.58
9,800个名字	9,800	6,972	100.00
25,856 姓名	25,856	21,051	100.00

表2：数据集统计信息。对于 Ngender 数据集，由于它是单字符数据集，我们将其唯一字符计为不同名字。此外，我们还展示了每个数据集中男性与女性的比例。

数据集

我们提供了一个数据集，其中包含58,393,173条记录，涵盖560,706个不同的名字，并记录了每次名字出现时对应的性别，数据收集自官方来源。具体而言，我们首先从中国国家工商行政管理总局的公司注册数据集中获取8,224,820个不含性别信息的不同全名。随后，我们将这些全名作为查询条件，向广东省政府的一项服务发起查询，该服务会提供查询姓名对应的女性和男性人数。最后，我们将性别频数按560,706个不同名字汇总，形成我们的数据集。作为对比，GitHub上最流行的中文姓名性别预测工具Ngender提供了一个包含32,067,566条记录和9,442个汉字的数据集，其数据收集自非官方来源。数据集的详细信息见表2。

与 Ngender 数据集相比，我们的数据集天然包含更多信息——Ngender 仅提供每个汉字出现在姓名中的特定位置时，具有该汉字的女性和男性人数，而缺少多字符组合的信息。然而，这种组合能够传达重要的性别信息。例如，正如引言中所讨论的，当两个与同一性别相关联的汉字组合在一起时，其组合可能指向相反的性别。根据我们的统计，这类情况占我们数据集中姓名的 1.75%。此外，我们还发现，将姓名中的两个汉字顺序调换后，14.77% 的姓名会呈现相反的性别倾向。这些结果证实，字符组合确实能为区分性别提供有用信息。

此外，我们的数据集性别分布更加均衡，男性与女性的比例（M-to-F）为 111.59%，而 Ngender 的比例则高达 197.28%，极不均衡。

实验

实验设置

数据集。 除了我们自己的数据集外，我们还使用前文提到的 Ngender 数据集训练模型，以比较数据质量提升带来的效果。我们不仅通过划分训练集和测试集来测试模型，还引入了来自两个独立来源的测试数据，其中包含姓名及其性别信息。我们分别将这两个数据集命名为 **9,800 个姓名**（Cai 等人（2021））和 **25,856 个姓名**（Du、Liu 和 Tian（2020））。其详细信息见表 2。

实现细节。 Ngender 数据集和我们的数据集均划分为 90% 训练集、5% 验证集，以及 5% 测试集。所有模型均使用相同的训练轮数进行训练。每个模型的学习率和权重衰减值均通过网格搜索进行调整。我们使用准确率来评估所有模型，即正确猜测的样本数占总样本数的比例。

所有模型中的初始嵌入均为随机初始化。对于所有任务，我们将注意力头数量设为 6，嵌入向量维度设为 768。我们使用 AdamW 作为优化器。所有模型的学习率和权重衰减均通过网格搜索进行调整。

基线方法。 我们将自己的方法（CHGAT）与三种有代表性的基线方法进行比较:

- **Ngender：** 一种常用的基于朴素贝叶斯的中文姓名性别预测工具。
- **拼音BERT（Jia 和 Zhao，2019年）：** 拼音BERT（PBERT）利用预训练 BERT 中的汉字语义，以及读音所承载的性别信息，将两种嵌入拼接后输入 BERT 模型。
- **FGAT（Wang 等人（2021年））：** 汉字构形图注意力网络是汉字表示学习领域的先进模型。它是一种多任务表示模型，利用同构图捕捉汉字部件所承载的语义信息。为进行公平比较，我们将姓名的发音嵌入与 FGAT 输出的字符嵌入拼接，以纳入发音信息，并使用拼接后的嵌入预测姓名性别。

实验结果

表 3 显示，我们的模型在所有训练与测试组合中都取得了最高准确率。当实验完全在我们的数据集上进行时，准确率达到 93.62%，显著高于 Ngender 在其公开数据集上的表现（84.76%）。

作为基于图的方法，我们的方法和 FGAT 在所有实验中的表现都优于 PBERT，相对提升最高达 4.72%。这表明所捕获的结构信息在姓名性别推断中发挥着重要作用。我们的模型进一步超越了当前最先进水平 FGAT：在 Ngender 数据集上训练时，准确率高出 0.14% 至 1.27%。这说明部件关系的异质性以及共享发音所传达的信息有助于预测。

值得注意的是，在我们的数据集上训练的模型，其表现均优于在 Ngender 数据集上训练的自身表现；测试集来自 Ngender 时除外（表 3 中的“Ngender”列）。这表明我们的数据集是更好的训练数据来源。出现例外的一个可能原因是，模型可能学到了依赖于特定数据集的信息，而这些信息有助于预测来自同一数据集的测试样本。

消融实验

为验证异构图中纳入发音节点类型的做法，以及三级注意力网络的设计考量，我们构建了网络的两个变体（变体1和变体2），并将它们与原始模型和FGAT进行比较。变体1通过移除

训练	方法	测试数据集			
		9,800 个姓名	25,856 个姓名	Ngender	Ours
Ours	Ngender	0.7066	0.7529	0.6636	0.8757
	PBERT	0.8136	0.8081	0.7849	0.9309
	FGAT	0.8139	0.8126	0.7854	0.9329
	CHGAT	**0.8147**	**0.8186**	**0.7873**	**0.9362**
Ngender	Ngender	0.6868	0.7518	0.6636	0.8476
	PBERT	0.7541	0.7776	0.8010	0.9012
	FGAT	0.7798	0.7977	0.8042	0.9148
	CHGAT	**0.7897**	**0.8040**	**0.8054**	**0.9168**

表3：在Ngender数据和我们的数据上训练，并在四个数据集上测试的所有模型的实验结果。

原始模型CHGAT层中的发音节点类型和汉字发音-字符元路径图得到（详情见图 4 上半部分）。用于聚合发音和结构信息的聚合注意力层不再需要，因此将其移除。在变体2中（见图4下半部分），我们移除结构注意力层，并直接使用聚合注意力层。

在**9,800 个姓名**数据集上（按8:1:1划分为训练集、验证集和测试集），如表4所示，两个变体的表现均低于原始模型，但仍优于FGAT模型。变体1的表现比FGAT相对提升了1.27%，这表明，与FGAT中的同构图相比，变体1中的异构图显然能够捕获更多结构信息。

不出所料，我们的模型表现比变体2相对提升了1.26%，这表明原始模型中的多层级注意力网络比单层注意力网络更有效。

尽管变体2包含发音信息，但其表现不如变体1。这表明，变体2的单层注意力在整合发音信息方面效果很差，甚至会引入噪声，损害其表现。

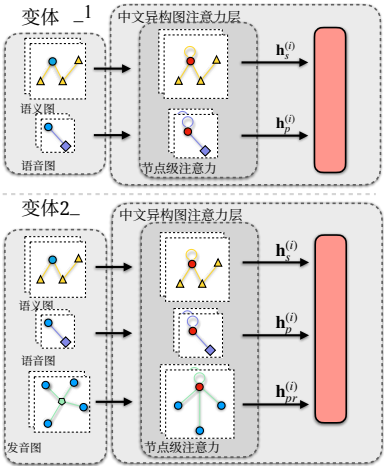

图4：CHGAT层各变体示意图。

	FGAT	变体1 _	变体2 _	我们的模型
准确率	0.8010	0.8112	0.8071	0.8173

表4：FGAT、变体1、变体2和我们的模型在**9,800 个姓名**上训练和测试所得的准确率。

复杂度分析

我们分析基线方法和我们模型的复杂度。对于 PBERT，复杂度为 $O(dn^2)$，其中 d 表示特征维度，n 是序列长度（姓名中拼音字母的数量）。FGAT包含一个词图学习部分，会使复杂度增加至 $O(dn^2 + L|V|d^2 + L|E|d)$，其中 L 是GNN层的数量，$|V|$ 是节点数量，$|E|$ 是边的数量。与 FGAT 相比，我们的模型还会执行额外的注意力聚合，使复杂度增加 adp^2，其中 p 表示元路径的数量，a 是聚合次数。这一增幅很小，因为复杂度主要由 d^2决定。此外，由于训练在线下进行，且在猜测性别时不需要即时响应，因此该复杂度在实际应用中可以接受。

结论

为了解决高质量中文姓名性别预测工具匮乏的问题，并促进针对代表性不足群体的性别偏见研究，我们提出了一种用于中文姓名性别预测的异构图注意力模型，该模型融合了汉字的结构和发音信息，性能优于所有最先进模型。此外，我们开源了一个大规模中文姓名性别数据集以及源代码。未来，我们计划将该方法扩展到许多其他涉及汉字表示的任务中。

致谢

本项目部分经费由上海市科学技术委员会上海市软科学研究计划重点项目资助（编号：22692112900）。

参考文献

Cai, H.; Jing, Y.; Wang, J.; 等. 2021. 中国独特姓名日益普遍的新证据：对 Ogihara 的回应。心理学前沿，12。Cao, S.; Lu, W.; Zhou, J.; and Li, X. 2018. cw2vec：利用笔画 n-gram 信息学习中文词嵌入。载于第三十二届 *AAAI*人工智能会议。Cimpian, J. R.; Kim, T. H.; and McDermott, Z. T. 2020. 理解 STEM 领域持续存在的性别差距。科学，368(6497)：1317–1319。Du, B.; Liu, P.; and Tian, Y. 2020. 百年来中文姓名性别特征的量化研究。载于第十九届中国计算语言学大会论文集，20–30。Hu, Y.; Hu, C.; Tran, T.; Kasturi, T.; Joseph, E.; and Gilling- ham, M. 2021. 姓名意味着什么？——基于字符的机器学习模型进行姓名性别分类。数据挖掘与知识发现，1–27。Huang, J.; Gates, A. J.; Sinatra, R.; and Barab´asi, A.-L. 2020. 跨国家和学科对科学职业中性别不平等的历史比较。美国国家科学院院刊，117(9)：4609–4616。Jensen, K.; Kov´acs, B.; and Sorenson, O. 2018. 获取和维持专利权方面的性别差异。自然生物技术，36(4)：307–309。Jia, J.; and Zhao, Q. 2019. 基于中文姓名的性别预测。载于 *CCF*自然语言处理与中文计算国际会议，676–683。Koning, R.; Samila, S.; and Ferguson, J.-P. 2021. 我们为谁发明？女性申请的专利更关注女性健康，但女性发明者寥寥无几。科学，372(6548)：1345–1348。Larivi`ere, V.; Ni, C.; Gingras, Y.; Cronin, B.; and Sugimoto, C. R. 2013. 文献计量学：科学领域的全球性别差异。自然，504(7479)：211–213。Lauscher, A.; Glava˘s, G.; Ponzetto, S. P.; and Vuli´c, I. 2020. 分布式词向量空间中隐式与显式去偏的通用框架。载于 *AAAI*人工智能会议论文集，8131–8138。Matthews, S.; Hudzina, J.; and Sepehr, D. 2022. 法律意见文本词嵌入中的性别与种族刻板印象检测。*AAAI*人工智能会议论文集，36(11)：12026–12033。Mukherjee, S.; and Bala, P. K. 2017. 基于作者风格的微博文本性别分类。信息系统与电子商务管理，15(1)：117–138。Ritter-Hayashi, D.; Vermeulen, P.; and Knoben, J. 2019. 这是一个男性主导的世界吗？性别多样性和性别平等对企业创新能力的影响。*PLOS ONE*，14(9)：1–19。Sun, J. 1997.《现代形声字辨析》。文科教学，(2)：58–65。Sun, Y.; Lin, L.; Yang, N.; Ji, Z.; and Wang, X. 2014. 增强偏旁信息的中文字符嵌入。载于国际神经信息处理会议，279–286。Tao, H.; Tong, S.; Zhao, H.; Xu, T.; Jin, B.; and Liu, Q. 2019年. 部首感知注意力模型，用于中文文本分类. In *AAAI*人工智能会议论文集, 5125–5132. van den Oever, K. F. 2021. 匹配中层与高层管理者：中层与高层管理者之间的性别和任职年限相似性是否会影响组织绩效？*PLOSONE*, 16(3): 1–18. Wais, K. 2016. 基于名字的性别预测方法及其 genderizeR 应用.《*R*期刊》, 8(1): 17. Wang, X.; Ji, H.; Shi, C.; Wang, B.; Ye, Y.; Cui, P.; and Yu, P. S. 2019年. 异构图注意力网络. In 万维网大会, 2022–2032. Wang, X.; Xiong, Y.; Niu, H.; Yue, J.; Zhu, Y.; and Yu, P. S. 2021. 利用构形图注意力网络改进汉字表示. In第*30*届*ACM*信息与知识管理国际会议论文集, 1999–2009. Wu, C.; Wu, F.; Liu, J.; He, S.; Huang, Y.; and Xie, X. 2019年. 利用搜索查询进行神经人口统计预测. In 第十二届*ACM*网页搜索与数据挖掘国际会议论文集, 654–662. Yang, Z.; and Feng, J. 2020. 一种减少词嵌入关系中性别偏见的因果推断方法. In*AAAI*人工智能会议论文集, 9434–9441. Yin, R.; Wang, Q.; Li, P.; Li, R.; and Wang, B. 2016. 多粒度中文词语嵌入. In *2016*年自然语言处理实证方法会议论文集, 981–986. Yu, J.; Jian, X.; Xin, H.; and Song, Y. 2017. 中文词语、汉字及细粒度汉字部件的联合嵌入. In *2017*年自然语言处理实证方法会议论文集, 286–291. Zhang, Y.; Liu, Y.; Zhu, J.; Zheng, Z.; Liu, X.; Wang, W.; Chen, Z.; and Zhai, S. 2019年. 从汉字的笔画、结构和拼音中学习中文词语嵌入. In第*28*届*ACM*信息与知识管理国际会议论文集, 1011–1020. Zhao, H.; and Kamareddine, F. 2017. 用于名字的高级性别预测工具及其在分析英国、马来西亚和中国计算机科学领域性别差异中的应用. In *2017*年计算科学与计算智能国际会议, 222–227.