For the Underrepresented in Gender Bias Research: Chinese Name Gender Prediction with Heterogeneous Graph Attention Network

Zihao Pan*, Kai Peng*, Shuai Ling, Haipeng Zhang †

ShanghaiTech University
{panzh, pengkai, lingshuai, zhanghp}@shanghaitech.edu.cn

Abstract

Achieving gender equality is an important pillar for humankind's sustainable future. Pioneering data-driven gender bias research is based on large-scale public records such as scientific papers, patents, and company registrations, covering female researchers, inventors and entrepreneurs, and so on. Since gender information is often missing in relevant datasets, studies rely on tools to infer genders from names. However, available open-sourced Chinese gender-guessing tools are not yet suitable for scientific purposes, which may be partially responsible for female Chinese being underrepresented in mainstream gender bias research and affect their universality. Specifically, these tools focus on character-level information while overlooking the fact that the combinations of Chinese characters in multi-character names, as well as the components and pronunciations of characters, convey important messages. As a first effort, we design a Chinese Heterogeneous Graph Attention (CHGAT) model to capture the heterogeneity in component relationships and incorporate the pronunciations of characters. Our model largely surpasses current tools and also outperforms the state-of-the-art algorithm. Last but not least, the most popular Chinese name-gender dataset is single-character based with far less female coverage from an unreliable source, naturally hindering relevant studies. We open-source a more balanced multi-character dataset from an official source together with our code, hoping to help future research promoting gender equality.

Introduction

Recently, there have been increasing gender-equality studies, regarding female researchers (Larivière et al. 2013; Huang et al. 2020), inventors (Jensen, Kovács, and Sorenson 2018; Koning, Samila, and Ferguson 2021), entrepreneurs (Ritter-Hayashi, Vermeulen, and Knoben 2019; van den Oever 2021), and STEM students (Cimpian, Kim, and McDermott 2020), based on large-scale public records such as scientific papers, patents, and company registrations. Given the fact that many of these datasets do not contain gender information, genders are usually inferred from individual names. Surprisingly, Chinese females are underrepresented in such research, though there is abundant comparable data in Chinese. One possible reason is the lack of reliable Chinese name-gender guessing tools that suit the standard of scientific purposes, compared with their English counterparts.

Ngender[1] is a basic tool for Chinese name-gender guessing, based on Naïve Bayes. It calculates the probability of a first name, often consisting of one or two Chinese characters, being a female name, by multiplying such probabilities of individual characters. Though straightforward, Ngender has two limitations. One is associated with Naïve Bayes itself – it does not work for out-of-sample characters that the classifier has never seen. Furthermore, it overlooks the knowledge from the combination of characters as well as the character components. The Ngender training data also deepens both limitations – it only contains the numbers of times each available character appears in names of females and names of males, instead of the frequencies of complete first names under both genders. Existing studies have proved that word representations from neural network language models such as BERT and GloVe have a gender tendency and convey gender information (Jia and Zhao 2019; Yang and Feng 2020; Lauscher et al. 2020; Matthews, Hudzina, and Sepehr 2022). Among them, Jia and Zhao (2019) propose a BERT-based model in which each character, whether in or out of the sample, gets a representation from a pre-trained BERT that handles the 'character-out-of-sample' problem for name-gender prediction. Besides, by concatenating the character embedding with the pronunciation embedding, their model, Pinyin BERT (PBERT), also proves that pronunciations deliver gender information.

However, beyond semantics from characters themselves and their pronunciations, the semantics arisen from character components are overlooked in current gender guessing tools. A large portion of Chinese characters consists of components, and these components help shape the meanings of the characters (Yu et al. 2017). Following this lead, studies utilize the component-level internal semantic features for Chinese character representation learning in a word2vec fashion (Sun et al. 2014; Yu et al. 2017; Zhang et al. 2019). In this direction, Wang et al. (2021) takes a step forward, capturing the semantic relationships between char-

*These authors contributed equally.

†*Corresponding author.

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

acters with shared components by constructing a homogeneous graph. In this way, characters and their components are inter-linked such that the semantic relationships between characters are better shaped and weighted with the attention mechanism. Though their FGAT model is the SOTA for many downstream NLP tasks, the relationships it relies on are in fact often heterogeneous, and it has not considered the same-pronunciation connections which can potentially augment the graph, as hinted by PBERT.

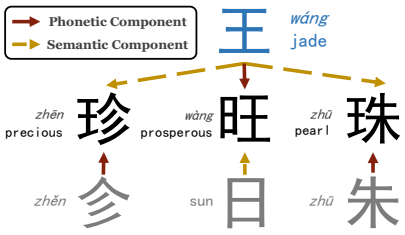

Figure 1: Example of the shared-component connections. '王' is the semantic component of '珍' and '珠', but is the phonetic component of '旺'.

We illustrate the heterogeneity in shared-component connections with an example and discuss how it affects our model design. In Figure 1, we see that the character '珍' (precious, *zhēn*)[2], '珠' (pearl, *zhū*) and '旺' (prosperous, *wàng*) share the component '王' (jade, *wáng*). If modeled by a homogeneous graph as in FGAT, it will indicate that '珍', '珠', and '旺' have equal pair-wised semantic similarity, contributed by the shared component '王'. Actually, '珍' and '珠' are much closer semantically, than to '旺'. Regarding genders, '珍' and '珠' are popular in female names, while '旺' has a strong male tendency. Therefore, modeling their relationships homogeneously would mislead name-gender prediction. In fact, their relationships can be distinguished if we know the '王' in '旺' is a phonetic component (solid arrow in Figure 1) which indicates the pronunciation of '旺' and the '王' in '珍' and '珠' is a semantic component (dashed arrow in Figure 1) contributing to its meaning. Besides '王', '旺' also has a semantic component '日' (sun) as shown at the bottom of Figure 1. Similarly, '珍' and '珠' has '彡' (*zhěn*) and '朱' (*zhū*) as their phonetic components, respectively. A character with both semantic component and phonetic component, such as '珍', '珠' and '旺', is called a picto-phonetic character. 80.5% of Chinese characters are picto-phonetic (Sun 1997), suggesting the effect of the heterogeneity in shared-component relationships is non-negligible. Therefore, we design heterogeneous graphs that specify character-semantic component edges and character-phonetic component edges.

Within this model structure, we introduce the shared-pronunciation connection as a new type of edges, given the effectiveness of pronunciations in gender guessing. As a straightforward example, characters sharing the pronunciation 'mei' (the same pronunciation of the character '美', which means beautiful) are more likely to be in female names. Now with character-semantic component edges, character-phonetic component edges, and character-pronunciation edges, we use a multi-level Chinese character attention network first to learn the importance of different components and structural information at the component-level and then aggregates the gender information conveyed by pronunciations (i.e., pinyin).

In addition to methodological contributions, we provide a high-quality Chinese name-gender dataset. Most Chinese name-gender datasets, such as the Ngender dataset, only contains frequencies of individual characters instead of complete names and loses important information for Chinese name-gender prediction. Unlike English first names, which are usually one-word names, most Chinese first names have one or two characters, with two-character names being the majority (84.55%)[3]. For Chinese first names with more than one character, the combinations of characters can be informative, and they sometimes even deliver opposite information as we can get from individual characters. For instance, '胜'(win) and '男'(male) are characters appearing more in male names, but '胜 男'(triumph over males) is a female name. Furthermore, character combinations A-B and B-A sometimes differ in gender probabilities, as suggested by our analysis in the Dataset section. To unleash the power of sequential representations and promote relevant research, we open source our large-scale full first name data, collected from an official source.

To sum up, our contributions in this paper include:

1. To the best of our knowledge, this is the first work that uses a graph neural network to model Chinese characters' internal and external connections for name-gender prediction to facilitate gender-equality studies.
2. We propose a heterogeneous graph with a multi-level attention network to capture the heterogeneity in semantic relationships between characters and components, as well as gender inclination indicated by pronunciations. The model outperforms various baselines and reaches a state-of-the-art accuracy of 93.62%.
3. We provide a dataset of 58 million Chinese names with associated genders as well as our source code[4], in hopes of promoting gender equality research, especially for the underrepresented Chinese females.

Related Work

Name-gender Prediction

Name-gender prediction is a common task in gender-quality studies (Larivière et al. 2013; Huang et al. 2020; Koning, Samila, and Ferguson 2021), as well as in web-based services like advertising and recommendation systems (Mukherjee and Bala 2017; Wu et al. 2019). Names, being very informative in many languages and cultures, are one important clue for gender guessing.

For western names, Wais (2016) designs genderizeR to predict the gender of an input name, simply according to the

[2]The meaning and the pronunciation (italic) are in parentheses.

[3]www.mps.gov.cn/n2253534/n2253535/c8349222/content.html

[4]https://github.com/ZhangDataLab/CHGAT

majority gender of people under this name in their data. Neural network language models further help exploit more information and tackle the 'out-of-sample' problem. Hu et al. (2021) construct a character-level BERT-based model to guess genders from English names.

Chinese, different from Latin-based languages, is logo graphic. Besides the Chinese characters, their components and pronunciations also convey information (Cao et al. 2018). Furthermore, in multi-character Chinese names, combinations across characters bring additional information. However, most gender-guessing tools only rely on character-level information. For instance, Ngender and the model proposed by Zhao and Kamareddine (2017) both regard the product of all characters' probabilities under a gender as the name's probability of being that gender, and output the gender with the highest probability. They have natural limitations in capturing the extra information that comes with the combinations of characters. Jia and Zhao (2019) concatenate character embeddings and pronunciation embeddings from pre-trained BERT models and mitigate these limitations. However, the connections between Chinese characters indicated by shared components and pronunciations are yet to be fully exploited.

Representing Characters with Components

The components of Chinese characters convey rich semantic information. Previous work incorporates the component information into the character embeddings based on the word2vec model to learn the representation of the Chinese characters (Sun et al. 2014; Yin et al. 2016; Yu et al. 2017). Seeking finer-grained information, some studies break components into sequenced strokes (subcomponents) to enhance the representations (Cao et al. 2018; Zhang et al. 2019). Though the characters are decomposed into components or subcomponents, the word2vec model cannot discriminate their importance and irrelevant parts can introduce noises into the representations.

Hence, Wang et al. (2021) model a character and its components in a homogenous graph with attention to learn the importance of components. Meanwhile, it shapes the character semantics through the characters' connections with other characters sharing the same components. As a result, their FGAT model achieves SOTA results on various downstream NLP tasks. However, as discussed in Introduction, their homogenous graph cannot model the heterogeneity in character-component relationships, since the majority of Chinese characters are picto-phonetic and the character-semantic component relationships and character-phonetic component relationships should be specified individually in the form of heterogeneous graphs. As an add-on, the shared-pronunciation relationships can also be integrated in heterogeneous graphs, potentially bringing more gender information for our prediction task.

Method

The overall structure of our model is shown in Figure 2. We take the Chinese characters and their pronunciations as the inputs to learn the information from the intra-character and intra-pronunciation combinations in names. Input names in the form of Chinese characters go into a Chinese Heterogeneous Graph Attention (CHGAT) layer and the BERT text encoder layer simultaneously. The output embeddings from the two layers are added and concatenated with the name's pronunciation embedding, generated by the BERT text encoder layer. This embedding is then fed into the Transformer encoder module to learn the contextual information within names. Finally, the classifier, a single layer fully connected network, is used to predict the gender. We detail the design of our heterogeneous graph as well as its core component, the CHGAT layer, in the following sections.

Heterogeneous Graph Structure

Formation of Chinese Characters. Chinese evolves from an ancient hieroglyph writing system. Basic characters such as wood and fire appeared first, with their individual graphical representations. Characters indicating more complex concepts, are composed by the combinations of these basic characters (Tao et al. 2019).

For example, by paralleling two '木 (wood, *mū*)' we get '林 (woods, *lín*)'. When '林 (woods, *lín*)' is further combined with '火 (fire, *huǒ*)' underneath it, it becomes '焚 (burn, *fén*)'.

Although the shapes of characters change over time, how they are combined (i.e., their structures) are preserved. There are 17 types of structures in modern Chinese according to QXK[5] (a Chinese character information system built by researchers from Beijing Normal University) and we list them in Table 1. As mentioned in Introduction, 80.5% of simplified Chinese characters are picto-phonetic, and QXK provides the semantic components and phonetic components of most Chinese characters.

Graph Structure. In this paper, we define a heterogeneous graph with four types of nodes, namely, ***character***, ***semantic component***, ***phonetic component***, and ***pronunciation***, and three types of paths, including ***character-semantic component***, ***character-phonetic component***, and ***character-pronunciation***, to incorporate gender information.

If a character is picto-phonetic, it connects with its one component representing the sound through a character-phonetic component edge. For its other components, it connects with them by character-semantic component edges. If a character is non-picto-phonetic, it connects with its components through character-semantic component edges as well. Besides, a character-pronunciation edge is added between the character and its pronunciation.

We take an example of the character '珠' to demonstrate how characters, pronunciations, and different types of components are represented and connected in our heterogeneous graph. In Figure 3, the focal character '珠' and its phonetic component '朱' are connected by a character-phonetic component edge, and character '珠' and its semantic component '王' are connected by a character-semantic component edge.

Previous research (Wang et al. 2021) utilizes the homogeneous graph to formulate the relations of components to characters, where there is only one type of node representing all components and characters and one type of edge

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

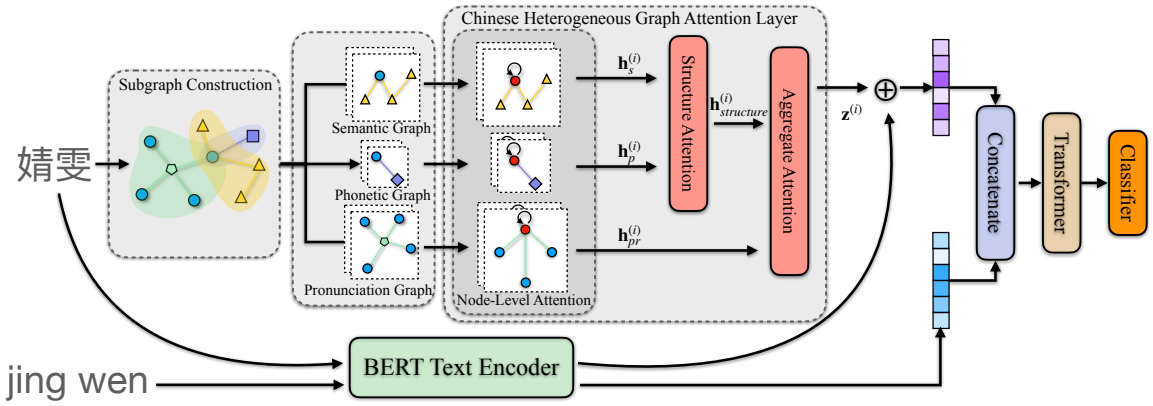

Figure 2: System structure.

Index	Structure Type	Icon	Example
1	left to right	⿰	桦, 婧
2	left to middle and right	⿲	树, 辫
3	above to below	⿱	雯, 芳
4	above to middle and below	⿳	衷, 害
5	full surround	⿴	园, 围
6	surround from above	⿵	阔, 闯
7	surround from below	⿶	凶, 函
8	surround from left	⿷	区, 匠
9	surround from upper left	⿸	房, 庐
10	surround from upper right	⿹	勾, 氦
11	surround from lower left	⿺	赶, 迁
12	integral	□	一, 口
13	isosceles triangle layout	△	鑫, 森
14	square layout	⊕	品, 燚
15	multielement combination	◎	彝, 燮
16	overlaid	§	爨, 南
17	multielement stacking	※	爽, 坐

Table 1: The list of Chinese characters' formation types. The first 11 structure type names follow their unicode names, while the last 6 have no formal names, so we describe their structural characteristics as best we can.

for inter-character shared-component relationships. They assume that a glyph represents the same semantics when used as a character, a semantic component, or a phonetic component, which is oversimplified. We distinguish between the semantic and phonetic components, and connect the focal character to all the characters sharing the same components.

As discussed in Introduction section, Jia and Zhao (2019) show that character pronunciations have gender inclinations, but they only use pronunciations as independent representations, without the message interacting with character embeddings. Hence, we connect the character '珠' and its pronunciation '*zhū*' by a character-pronunciation edge, as shown in Figure 3.

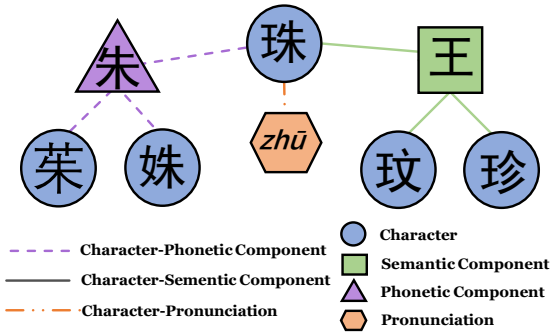

Figure 3: An example of graph composition of the character '珠'. '朱', 'zhū' and '王' are the phonetic component, pronunciation and semantic compoente of '珠', and they are connected with '珠' through the character-phonetic component, character pronunciation and character-sementic component, respectively.

Chinese Heterogeneous Graph Attention

To aggregate the information in the heterogeneous graph, we design the Chinese Heterogeneous Graph Attention (CHGAT) layer with three-level attentions. Both the semantic graph and the phonetic graph contain structural information. After the node-level attention aggregates the information within each graph, structural information remains scattered among *semantic component* and *phonetic component* nodes. Therefore, the network captures the structural information of a character by aggregating the two node-level representations. To assemble information from the pronunciation and the structure representations, another attention, i.e., the aggregate attention, is used.

We denote the set of Chinese characters as $\mathcal{C} = \{c_0, c_1, ..., c_m\}$. A character c_i's feature embedding is $\boldsymbol{f}_c^{(i)}$. All characters share one pronunciation graph $\boldsymbol{g}_{pr}$, which is a 'character-pronunciation-character' meta-path in (Wang et al. 2019)'s definition, but each character has its own semantic graph and phonetic graph. For a character c_i, its semantic graph contains all the one-hop and two-hop semantic components of c_i, denoted as $\boldsymbol{g}_s^{(i)}$, while its phonetic graph

is the one containing the one-hop phonetic component of c_i, represented by $\boldsymbol{g}_p^{(i)}$. The feature embedding of component s_o in the semantic graph is $\boldsymbol{f}_s^{(o)}$, while the feature embedding of component p_t in the phonetic graph is $\boldsymbol{f}_p^{(t)}$.

Wang et al. (2021) introduce position embedding that adds position information to the component representations. The same components may appear in various positions of different characters, and this position information can affect the meanings of these components. Hence, we add another position embedding, which represents the position of each component in the character, denoted as

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

Therefore, the initial embedding $\boldsymbol{x}_c^i$, the semantic components' initial embedding $\boldsymbol{x}_s^o$, and the phonetic components' initial embeddings $\boldsymbol{x}_p^t$ for an input character are:

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

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

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

respectively. i, o, and t denote the index of character, semantic component, and phonetic component, respectively. l_c, l_s, and l_p represent their corresponding position index, respectively.

Here, we define character c_i's phonetic graph feature embeddings as $\boldsymbol{X}_p^{(i)}$, and its semantic graph feature embeddings as $\boldsymbol{X}_s^{(i)}$. The pronunciation graph feature embedding is $\boldsymbol{X}_{pr}^{(i)}$.

Inspired by HAN (Wang et al. 2019), we use a three-level attention mechanism in our scenario. As illustrated in Figure 2. The model first learns the node-level embeddings within each graph, and then the node-level embeddings of a semantic graph and phonetic graph are fed into the structure attention to learn a structure embedding. This embedding and the node-level pronunciation embedding are used to learn the character's final embedding with an aggregated attention.

Node-level Attention. We first project different types of node features into the same feature space with a transformation matrix:

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

where $\boldsymbol{W}_k^{(i)} \in \mathbb{R}^{d_k \times d}$ is a learnable parameter. $k \in \{s, p, pr\}$ represents path type, and d_k is the input feature dimension. The importance score of node j to target node i is computed as:

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

where $\boldsymbol{w}_k \in \mathbb{R}^{1 \times 2d}$ is a learnable vector. The importance score is then normalized to be the weight of node j to node i, denoted as $\theta_k^{(i,j)}$:

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

where $\mathcal{N}_k^{(i)}$ is the set of all i's neighbors in the k type of path. The node-level embedding is then the weighted sum of all nodes connecting to itself:

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

where t is the number of heads, and $\|$ represents the concatenation operation.

Attention Module. The attention module aggregates the representation with different semantics into one representation. We denote it as:

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

where v represents the number of inputs. The importance of each input to the target embedding is:

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

where $\boldsymbol{q}$, $\boldsymbol{W}$ and $\boldsymbol{b}$ are learnable parameters in the model. $r \in \{1, 2, ..., v\}$ is the type of semantics, and N is the input name. Then the importance score of each input is:

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

The target embedding is:

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

Structure Attention Layer. In the structure attention layer, we learn the structure representation of c_i from its node-level embeddings of the semantic graph and the phonetic graph with an attention module:

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

The $\boldsymbol{h}_{structure}^{(i)}$ denotes the structure representation of character c_i.

Aggregate Attention Layer. The aggregate attention layer assembles the character c_i' pronunciation representation, and its structure representation into one embedding. Again, this is achieved by applying an attention module, which is:

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

Finally, $\boldsymbol{z}^{(i)}$ is the output of c_i at the CHGAT layer.

Loss Function

Our objective function is defined as:

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

where J represents the number of data, y_j is the label of the j^{th} input, and $\hat{y}_j$ represents the predicted probability of y_j being the label.

	Records	Unique First Names	M-to-F%
Ngender	32,067,566	9,442	197.28
Our Dataset	58,393,173	560,706	111.58
9,800 Names	9,800	6,972	100.00
25,856 Names	25,856	21,051	100.00

Table 2: Statistics of the datasets. For the Ngender dataset, we count its unique characters as its unique first names, since it is a single-character dataset. Besides, we show the ratio of males to females for each dataset.

Dataset

We provide a dataset with 58,393,173 records of 560,706 different first names and the associated gender for each name occurrence, collected from an official source. To be specific, we begin with 8,224,820 unique full names without gender information in a company registration dataset from China's State Administration for Industry and Commerce. These full names are then used as queries to a service from Guangdong province government that provides the number of females and males with the querying name. The resulting gender frequencies are aggregated under the 560,706 unique first names to form our dataset.As a comparison, the most popular Chinese name-gender prediction tool (Ngender) on GitHub provides a dataset of 32,067,566 entries of 9,442 characters which is collected from unofficial sources. The detail information of the datasets is shown in Table 2.

Our dataset is naturally more informative than the Ngender dataset – Ngender only provides for each character the numbers of females and males with this particular character in their first names, and the information from multi-character combinations is absent. However, this combination conveys important gender information. For instance, when two characters associated with the same gender are put together, their combination can indicate the opposite gender, as discussed in Introduction. According to our statistics, these cases account for 1.75% of names in our dataset. Beyond this, we discover that when we reverse the two characters in names, 14.77% of names would have reversed gender tendency. These confirm that character combinations actually deliver helpful information for distinguishing genders.

Moreover, our dataset is more balanced with a male-to-female ratio (M-to-F) of 111.59%, against a highly unbalanced ratio of 197.28% for Ngender.

Experiments

Experimental Setup

Datasets. Besides our dataset, we use the aforementioned Ngender dataset to train the models as a comparison for enhancements from improved data quality. We test the models not only by splitting training and test sets, but also by introducing test data from two independent sources containing names and genders. We name them **9,800 Names** (Cai et al. 2021) and **25,856 Names** (Du, Liu, and Tian 2020) respectively. Their detailed information is in Table 2.

Implementation Details. Both the Ngender dataset and our dataset are split into 90% training, 5% validation, and 5% test. All models are trained with the same epochs. The learning rate and the weight decay value of each model are adjusted with grid search. We use the accuracy score to evaluate all models, which is the number of correctly guessed instances over the total instances.

The initial embeddings in all models are randomly initialized. For all tasks, we set the number of attention heads to 6 and the dimension of embedding vectors to 768. We use AdamW as the optimizer. The learning rate and the weight decay of all models are adjusted with grid search.

Baselines. We compare our method (CHGAT) with three representative baselines:

- **Ngender:** A commonly used Chinese name-gender prediction tool based on Naïve Bayes.
- **Pinyin BERT (Jia and Zhao 2019):** Pinyin BERT (PBERT) makes use of characters' semantics from a pre-trained BERT as well as the gender information delivered in pronunciations by concatenating the two embeddings and feeding them into a BERT model.
- **FGAT (Wang et al. 2021):** The Chinese character formation graph attention network is a state-of-the-art model in Chinese character representation learning. It is a multitask representation model that uses a homogeneous graph to capture the semantic information delivered by a character's components. To make a fair comparison, we include pronunciation by concatenating the name's pronunciation embedding to the output character embedding from FGAT, and use the concatenated embedding to predict name gender.

Experiment Results

Table 3 shows that our model hits the highest accuracy scores in all training and test combinations. When the experiment is conducted entirely on our dataset, it achieves the accuracy of 93.62% which is significantly higher than the public available Ngender on its public dataset (84.76%).

Our method and FGAT, as the graph-based methods, outperform PBERT in all experiments by up to 4.72% relatively. This indicates the structural information captured plays an important part in name-gender guessing. Our method further surpasses the current SOTA, FGAT, by 0.14% to 1.27% when trained on the Ngender dataset, suggesting the heterogeneity in component relationships and information conveyed in shared pronunciations help the prediction.

It is worth noting that models trained on our dataset all outperform themselves trained on the Ngender dataset, with exceptions when the test set is from Ngender (the 'Ngender' column in Table 3). This suggests that our dataset is a better source for training data and a possible reason for the exceptions is that dataset-dependant information may be learned, and it helps predict the test examples from the same dataset.

Ablation Study

To validate the pronunciation node type included in our heterogeneous graph and the design consideration of three-level attention network, we build two variants (variant_1 and variant_2) of our network and compare them with our original model and FGAT. variant_1 is obtained by removing

Training	Method	Testing Dataset			
		9,800 Names	25,856 Names	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**

Table 3: Experiment results of all models trained on the Ngender data and our data, and tested on four datasets.

the pronunciation node type and *character-pronunciation-character* meta-path graph from the CHGAT layer in the original model (details in the upper part of Figure 4). The aggregation attention layer used for aggregating pronunciation and structural information is no longer needed and therefore removed. In variant_2 (shown in the lower part of Figure 4), we remove the structure attention layer and use the aggregate attention layer directly.

On the **9,800 Names** dataset (split in 8:1:1 for training, validation, and test sets), the two variants show decreased performance from the original model but still remain more effective than the FGAT model as shown in Table 4. The variant_1 achieves a relative improvement of 1.27% compared to FGAT, which suggests that the heterogeneous graph in variant_1 obviously captures more structural information than the homogeneous one in FGAT.

Unsurprisingly, our model achieves a relative improvement of 1.26% compared to variant_2, which indicates that the multi-level attention network in the original model is more effective than the single-level attention network.

Though variant_2 includes pronunciation information, it has a worse performance compared with variant_1. This suggests that variant_2's single-level attention does a very bad job in incorporating pronunciation information, such that it even introduces noises that undermine its performance.

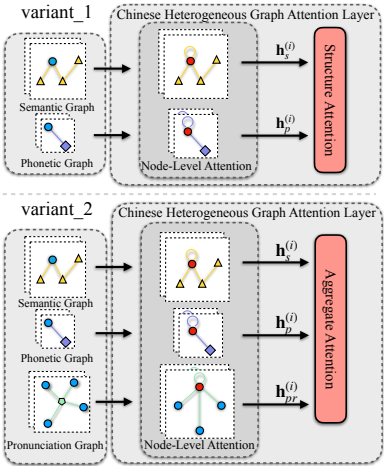

Figure 4: Illustration of CHGAT layer's variants.

	FGAT	variant_1	variant_2	our model
accuracy	0.8010	0.8112	0.8071	0.8173

Table 4: Accuracy of FGAT, variant_1, variant_2, and our model trained and tested on **9,800 Names**.

Complexity Analysis

We analyze the complexity of the baselines and our model. For PBERT, the complexity is $O(dn^2)$, where d denotes the feature dimension and n is the sequence length (number of Pinyin letters in a name). FGAT has a word graph learning part that increases the complexity to $O(dn^2 + L|V|d^2 + L|E|d)$, where L is the number of GNN layers, $|V|$ is the number of nodes and $|E|$ is the number of edges. Compared with FGAT, our model performs additional attention aggregations, adding adp^2 to the complexity, where p denotes the number of meta-paths and a is the number of aggregations. This is a slight increase since the complexity is largely determined by d^2. Besides, as the training happens offline and we do not need instant responses when guessing the genders, the complexity is acceptable in practice.

Conclusion

To address the lack of high-quality Chinese name-gender prediction tools and to facilitate the gender bias research for the underrepresented, we propose a heterogeneous graph attention model incorporating structural and pronunciation information of Chinese characters for Chinese name-gender prediction that outperforms all SOTA models. Besides, we open source a large-scale Chinese name-gender dataset as well as our source code. As a future step, we plan to extend our method to the many other tasks that involve Chinese character representations.

Acknowledgments

This project is partially supported by the Key Projects of Shanghai Soft Science Research Program, from the Science and Technology Commission of Shanghai Municipality (No. 22692112900).

References

Cai, H.; Jing, Y.; Wang, J.; et al. 2021. Novel evidence for the increasing prevalence of unique names in China: A reply to Ogihara. *Frontiers in Psychology*, 12.

Cao, S.; Lu, W.; Zhou, J.; and Li, X. 2018. cw2vec: Learning chinese word embeddings with stroke n-gram information. In *Thirty-second AAAI conference on artificial intelligence*.

Cimpian, J. R.; Kim, T. H.; and McDermott, Z. T. 2020. Understanding persistent gender gaps in STEM. *Science*, 368(6497): 1317–1319.

Du, B.; Liu, P.; and Tian, Y. 2020. A Quantified Research on Gender Characteristics of Chinese Names in A Century. In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, 20–30.

Hu, Y.; Hu, C.; Tran, T.; Kasturi, T.; Joseph, E.; and Gillingham, M. 2021. What's in a name?–gender classification of names with character based machine learning models. *Data Mining and Knowledge Discovery*, 1–27.

Huang, J.; Gates, A. J.; Sinatra, R.; and Barabási, A.-L. 2020. Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, 117(9): 4609–4616.

Jensen, K.; Kovács, B.; and Sorenson, O. 2018. Gender differences in obtaining and maintaining patent rights. *Nature Biotechnology*, 36(4): 307–309.

Jia, J.; and Zhao, Q. 2019. Gender prediction based on Chinese name. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 676–683.

Koning, R.; Samila, S.; and Ferguson, J.-P. 2021. Who do we invent for? Patents by women focus more on women's health, but few women get to invent. *Science*, 372(6548): 1345–1348.

Larivière, V.; Ni, C.; Gingras, Y.; Cronin, B.; and Sugimoto, C. R. 2013. Bibliometrics: Global gender disparities in science. *Nature*, 504(7479): 211–213.

Lauscher, A.; Glavaš, G.; Ponzetto, S. P.; and Vulić, I. 2020. A general framework for implicit and explicit debiasing of distributional word vector spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8131–8138.

Matthews, S.; Hudzina, J.; and Sepehr, D. 2022. Gender and Racial Stereotype Detection in Legal Opinion Word Embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11): 12026–12033.

Mukherjee, S.; and Bala, P. K. 2017. Gender classification of microblog text based on authorial style. *Information Systems and e-Business Management*, 15(1): 117–138.

Ritter-Hayashi, D.; Vermeulen, P.; and Knoben, J. 2019. Is this a man's world? The effect of gender diversity and gender equality on firm innovativeness. *PlOS ONE*, 14(9): 1–19.

Sun, J. 1997. Xian Dai Xing Sheng Zi Bian Xi (Modern Morpheme Analysis). *Wen Ke Jiao Xue*, (2): 58–65.

Sun, Y.; Lin, L.; Yang, N.; Ji, Z.; and Wang, X. 2014. Radical-enhanced chinese character embedding. In *International Conference on Neural Information Processing*, 279–286.

Tao, H.; Tong, S.; Zhao, H.; Xu, T.; Jin, B.; and Liu, Q. 2019. A radical-aware attention-based model for chinese text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5125–5132.

van den Oever, K. F. 2021. Matching middle and top managers: Do gender and tenure similarities between middle and top managers affect organizational performance? *PLOS ONE*, 16(3): 1–18.

Wais, K. 2016. Gender Prediction Methods Based on First Names with genderizeR. *The R Journal*, 8(1): 17.

Wang, X.; Ji, H.; Shi, C.; Wang, B.; Ye, Y.; Cui, P.; and Yu, P. S. 2019. Heterogeneous graph attention network. In *The World Wide Web Conference*, 2022–2032.

Wang, X.; Xiong, Y.; Niu, H.; Yue, J.; Zhu, Y.; and Yu, P. S. 2021. Improving chinese character representation with formation graph attention network. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 1999–2009.

Wu, C.; Wu, F.; Liu, J.; He, S.; Huang, Y.; and Xie, X. 2019. Neural demographic prediction using search query. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 654–662.

Yang, Z.; and Feng, J. 2020. A causal inference method for reducing gender bias in word embedding relations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 9434–9441.

Yin, R.; Wang, Q.; Li, P.; Li, R.; and Wang, B. 2016. Multi-granularity chinese word embedding. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 981–986.

Yu, J.; Jian, X.; Xin, H.; and Song, Y. 2017. Joint embeddings of chinese words, characters, and fine-grained subcharacter components. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, 286–291.

Zhang, Y.; Liu, Y.; Zhu, J.; Zheng, Z.; Liu, X.; Wang, W.; Chen, Z.; and Zhai, S. 2019. Learning Chinese word embeddings from stroke, structure and pinyin of characters. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1011–1020.

Zhao, H.; and Kamareddine, F. 2017. Advance gender prediction tool of first names and its use in analysing gender disparity in Computer Science in the UK, Malaysia and China. In *2017 International Conference on Computational Science and Computational Intelligence*, 222–227.