Gender Prediction Based on Chinese Name

Jizheng Jia[1] [0000−0002−8876−9644] and Qiyang Zhao[2] [0000−0002−8476−5742]

[1] Beihang University, Beijing, China
{zy1706128}@buaa.edu.cn
[2] Beihang University, Beijing, China
{zhaoqy}@buaa.edu.cn

Abstract. Much work has been done on the problem of gender prediction about English using the idea of probability models or traditional machine learning methods. Different from English or other alphabetic languages, Chinese characters are logosyllabic. Previous approaches work quite well for Indo-European languages in general and English in particular, however, their performance deteriorate in Asian languages such as Chinese, Japanese and Korean. In our work, we focus on Simplified Chinese characters and present a novel approach incorporating phonetic information (Pinyin) to enhance Chinese word embedding trained on BERT model. We compared our method with several previous methods, namely Naive Bayes, GBDT, and Random forest with word embedding via fastText as features. Quantitative and qualitative experiments demonstrate the superior of our model. The results show that we can achieve 93.45% test accuracy using our method. In addition, we have released two large-scale gender-labeled datasets (one with over one million first names and the other with over six million full names) used as a part of this study for the community.

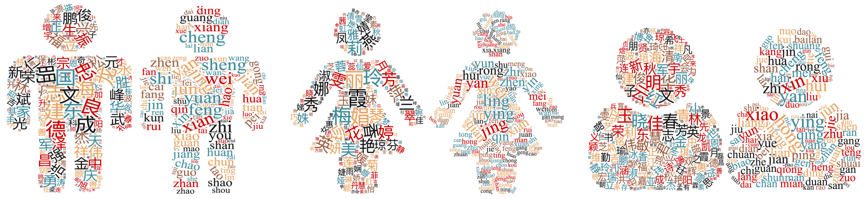
Keywords: Gender inference of names · Pinyin representation · BERT-based model.

1 Introduction

In recent years, the problem of classifying gender has occupied an important place in academia and industry. For many research questions, demographic information about individuals (such as names, gender, or ethnic background) is highly beneficial but it often particularly difficult to obtain [3,2].

The industry can also benefit from gender data and they can gain additional information about the customers, which would be used for targeted marketing or personalized advertisements [3,4]. Also, it can be used for name translation and automated gender selection in online form filling, making the software more convenient for users. Most Internet companies in China can access to National Identification Card (People's Republic of China) database, with an odd/even number indicating male/female with 100% accuracy. However, with increasing awareness of privacy data protection, more and more Internet users are reluctant to provide their personal data.

It has long been known that the appearance of a person is not independent of the name. Parents spend extraordinary time and effort to select the perfect name for an expected child. The name of a child conveys much more information including gender



(a) Male (Hanzi -> Pinyin) (b) Female (Hanzi -> Pinyin) (c) Unisex (Hanzi -> Pinyin)

Fig. 1: The Chinese character (Hanzi) of name can indicate some information of gender, in the same way, Pinyin is gender-specific. The characters of two figures (Hanzi and Pinyin) do not completely match because different characters that have the same pronunciation are reduced to the same Pinyin representation. For instance, in the Figure 1a, “中” and “忠” have the same Pinyin representation “zhong”.

both in western countries and China. As we all know, names can be divided into three parts: first, middle, and last name, which don’t have equal effects on gender. In mainland China, most people’s name consist of first name and last name. According to a report [23] on Chinese names from the Ministry of Public Security, there were 23 last names that claim more than 10 million users and the number of Chinese people bearing the last names Wang and Li has surpassed 100 million for both, which indicates million of people have the same last name. Maybe we can easily get the conclusion from the report that last name has little indication about gender information. In order to verify our inference, we provide two versions of name-gender datasets so as to compare the difference between the full name and first name in terms of gender prediction.

There have been several approaches proposed for gender estimation using first names. Most previous work were rules-based [5,16] or focused on probabilistic prediction using a large corpus of known names [7]. However, rule-based methods have drawbacks that the coverage of rules was limited and they are language dependent. Thus, it seems absurd to attempt to hard code every possible pattern when we have deep learning. What’s more, many previous work focused on alphabetic writing systems, their performances on Chinese deteriorated. So far, there are few studies on the Chinese language.

We intend to fill this gap by presenting a benchmark and comparison of several gender inference methods in the area of the Chinese language. There are many distinctions between Chinese and English languages. For example, Chinese sentences are not separated by spaces [17]. But the main distinction is that written Chinese is logosyllabic. A Chinese character has its own meaning or as a part of the polysyllabic word, making Chinese sentence segmentation a non-trivial task. In addition to the difficulty of sentence segmentation [9], most previous studies computed the semantic meaning of Chinese characters using the corpus from Baidu Baike[3] or Chinese Wikipedia[4]. This kind of method ignored the importance of pronunciations features. The work found that phonology also contributed to character recognition [10], which was ignored by most previous work.

[3] https://baike.baidu.com

[4] https://en.wikipedia.org/wiki/Main_Page

Thus, in our work, the Chinese word representations can be roughly divided into two components: semantic component and phonetic component. The semantic component indicates the meaning of a character while the latter indicates the sound of a character. Pinyin is the official phonetic representation for Chinese word, which converts Chinese character to Roman letters depending on the pronunciation of Mandarin Chinese [11]. For example, “qing” and “qin” are the phonetic component of characters “青” (seafoam) and “琴” (Qin, a Chinese musical instrument). In this case, they both contain female cues. Figure 1 shows that Pinyin is also gender inclination like Chinese characters (Hanzi). For example, in Figure 1a, “国” /guo/ and “德” /de/ are male inclined and in Figure 1c, we call “明” /ming/ and “晓” /xiao/ are unisex.

Then we used two representations: the simplified Chinese string and responding Pinyin representation. Our idea is that we can use phonetic features to enhance the original Chinese embeddings. Finally, we used the state-of-art model BERT [12] for gender classification based on the features we have described.

To summarize, the key contributions of this work are as follows: (1) This work fills the gap of integrating Pinyin information with traditional word embedding via BERT-Base model. (2) We compare several previous gender inference architectures with our model and the effectiveness of our idea is empirically well demonstrated. (3) To our best knowledge, this is the first large gender-labeled Chinese name dataset available for the community.

2 Related Work

Early work on gender prediction includes gender-guesser [22], Gender API [19] and Ngender [20]. The gender-guesser guesses gender from first English name using a dictionary. Sometimes, you get the “Unknown Token” if the input doesn’t exist in the dictionary. Gender API infers the likely gender of a name but it is a commercial application and not free for the community. Ngender also uses a predefined dictionary. The drawback is obvious: the dictionary is too limited when it comes to new words although this work[5] uses Laplacian smoothing technique addressing low-frequency words or OOV (out-of-vocabulary) problem.

More Recent studies pay more attention to traditional machine learning such as SVM [7,13] or boosted decision trees [5]. This work finds out a set of characteristics such as number of syllables, number of vowels and ending character, extracted from first name and then use these characteristics as feature to an SVM model [3].

In recent years, some effective methods to understand Chinese names have been proposed. This work provided a tool for both English and Chinese names simultaneously but only in STEM fields when they find no free gender forecasting software to predict global gender disparity [14]. [15] proposed a framework exploring both internal character information and external context of words to explore subword-level information of Chinese characters, however, in this paper, we propose to use Pinyin to enhance the traditional Chinese word embeddings.

In this paper, we’d like to use BERT model instead of traditional machine learning methods, which need manual feature extraction and selection. Moreover, we introduce

[5] http://sofasofa.io/tutorials/naive_bayes_classifier/

Pinyin as raw training data to original embedding hoping to enhance the ability of word embedding.

3 The Approaches and Data

3.1 Four Approaches

There are mainly four types of gender prediction models in our experiment: (1) a rule-based approach, which was omitted by its limitation in specific language or region [4], (2) probabilistic model like Naive Bayes, (3) two types of machine learning algorithms such as GBDT and Random forest, (4) BERT-Base model with character-Pinyin data.

Naive Bayes Classifier We firstly choose Naive Bayes for its simplicity since there is almost no hyper-parameter tuning needed. The Naive Bayes classifier uses Bayes Theorem to predict the probability of gender given name information [14]. The formula of Naive Bayes algorithm for gender prediction:

$$P(\frac{gender}{name}) = \frac{P(gender) \times P(\frac{gender}{name})}{P(name)}$$

In this formula, $P(gender|name)$ is the posterior probability of gender, given names; $P(name|gender)$ is the likelihood which is the probability of predictor, given gender; $P(name)$ is the prior probability of predictor.

We also apply Laplace smoothing when a frequency-based probability is zero, which is a practical technique used to deal with unknown words (names).

$$g(x) = \frac{n_x + \alpha}{l + \alpha c}$$

where n_x is the number of occurrences of the word x, l is the length of the sentence, c is the number of different words in the sentence, and α is the smooth coefficient of Laplacian smoothness, which is a hyper-parameter.

GBDT with Simple Features GBDT, however, is chosen as “simple version” of neural network and its decent performance in many machine learning competitions. The features are term frequency of every single Chinese character from the name.

Random Forest with Various Types of Features This model takes three types of features: the first name, the unigram of first name and a vector of size one hundred extracted from Skip-gram model via fastText trained separately using a corpus collected from Baidu Baike.

BERT-Base Model with Character-Pinyin Data The architecture of our model could be found in Figure 2. We mainly have to train the binary classifier, with some changes like the dictionary to the BERT-Base model during training process. The training process is also called Fine-Tuning.

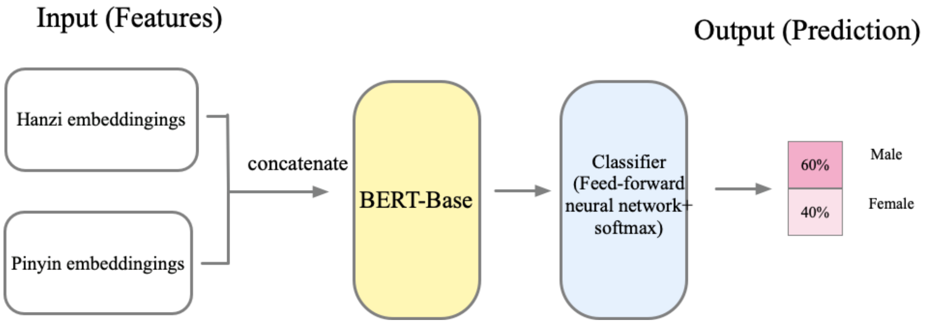


Fig. 2: BERT has achieved state-of-the-art results in wide variety of NLP language task. We load BERT-Base model, attach an additional layer for classification, and fine-tune BERT model for gender classification.

Data Type	Number	Percent
firstname (female)	718,400	63.70%
firstname (male)	409,411	36.30%
fullname (female)	4,366,278	66.25%
fullname (male)	2,223,964	33.75%

Table 1: The statistical information of our datasets. The difference between firstname dataset and fullname dataset is that the former has no first name. For example, “杨振宁” is a Chinese name that is officially displayed as Chenning Yang. And, in the fullname dataset, it’s showed as “yang chen ning” where chracters are separated by space and converted into lower case. But in the firstname dataset, it is represented as “chen ning”.

3.2 Gender-labeled Data

The gender-labeled data can be vital to gender classification. However, to the best of our knowledge, there is no suitable gender-labeled dataset for deep network training.

In our work, data was collected with reliable gender labels collected from several public sources. After cleaning and removing repetitive names, the final data lists consist of 6,590,242 names and 1,127,811 first names, respectively. We used pypinyin [21] and Google translator to translate those Chinese names into Pinyin format. Table 1 shows more details about our gender-labeled datasets.

There are two types of datasets: The first one is Simplified Chinese character which we call "hanzi". Since each Chinese character corresponds to a Pinyin, we take each Pinyin as a token corresponding to the Chinese character. Then we call the second type dataset as "character-pinyin".

4 Experiments

We explore four different classifier types described above. We choose Naive Bayes as baseline and GBDT using TF (term frequency) as the “simpler” version of neural

	Naive Bayes	GBDT	RF	Our Model
ACC (mean)	80.67%	85.59%	84.87%	**90.89%**
ACC (SD)	2.03	1.85	1.58	1.02

Table 2: Experiment results came from Chinese characters data without incorporating Pinyin information. We reported the mean and standard deviation (SD) of results from ten repeated experiments on each model with different random seed.

	Pinyin (only)	Hanzi (only)	Hanzi-Pinyin-first	Hanzi-Pinyin-full
ACC (mean)	69.64%	90.89%	**92.33%**	**93.45%**
ACC (SD)	1.78	0.96	1.02	1.52

Table 3: These experiments were trained on different combinations of our two datasets. Like last experiment, we used mean and SD of ACC to get unbiased result. ACC from "Hanzi-Pinyin-first" and "Hanzi-Pinyin-full" got better results than the others. And "Hanzi (only)" got better accuracy than "Pinyin (only)" when working alone.

network. The third model is based on Random forest using the word embedding trained by fastText. Finally, our model is trained on BERT using simplified Chinese words and corresponding Pinyin data.

Comparison Experiments were done on our first name dataset without Pinyin information. For fastText, we used Skip-gram to learn word embeddings, we set the minimum length of character n-grams to be 2 and the maximum length of character n-grams to be 4. We tuned the hyper-parameters using grid search and cross-validation with 90% in the training set, 5% in the dev set. The test data (5%) was held out for final comparison. The experimental results could be found in Table 2. The high accuracy in test set indicated the effiency of our proposed model.

To analyze the effect of Pinyin representation, we also did a second type of experiment applying different merging strategies of the data training on our model. The first type of dataset was first name data of Pinyin format; the second one was first name data of Simplified Chinese (Hanzi) format; the third one was first name data from both Pinyin and Simplified Chinese format, we called this as character-Pinyin first name (Hanzi-Pinyin-first); the final one used full name data from both Pinyin and Simplified Chinese format (Hanzi-Pinyin-full). Table 3 showed the detail information of our experiment results. You could find our codes and open-source datasets here[6] sooner.

5 Conclusion

Our experiments showed the efficiency of incorporating Pinyin information to enhance previous word embedding compared with using Chinese words only. We hold that this idea can also expand to other NLP tasks or downstream tasks.

Pinyin did not do as well as the Simplified Chinese characters, but combined with Simplified Chinese words, it can get better results compared to using Simplified Chinese

[6] https://github.com/jijeng/gender-prediction

words alone. Also, we face problem such as unisex names, which could be the direction for future work.

References

1. Juergen Mueller and Gerd Stumme. Gender inference using statistical name characteristics in twitter. In *Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on SocialInformatics 2016, Data Science 2016*, page 47. ACM, 2016.
2. Fariba Karimi, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 53–54, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.
3. Juergen Mueller and Gerd Stumme. Gender inference using statistical name characteristics in twitter. In *Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on SocialInformatics 2016, Data Science 2016*, page 47. ACM, 2016.
4. Monali Y Khachane. Gender estimation from first name: A rule based approach. *International Journal of Advanced Research in Computer Science*, 9(2):609, 2018.
5. Wendy Liu and Derek Ruths. What's in a name? using first names as features for gender inference in twitter. In *2013 AAAI Spring Symposium Series*, 2013.
6. Chuan Gu, Xi-ping Tian, and Jiang-de Yu. Automatic recognition of chinese personal name using conditional random fields and knowledge base. *Mathematical Problems in Engineering*, 2015, 2015.
7. John D. Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1301–1309, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
8. Ming Liu, Vasile Rus, Qiang Liao, and Li Liu. Encoding and ranking similar chinese characters. *J. Inf. Sci. Eng.*, 33(5):1195–1211, 2017.
9. Shilei Huang and Jiangqin Wu. A pragmatic approach for classical chinese word segmentation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018.
10. Nanyun Peng, Mo Yu, and Mark Dredze. An empirical study of chinese name matching and applications. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 377–383, 2015.
11. Yafang Huang and Hai Zhao. Chinese pinyin aided IME, input what you have not keystroked yet. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2923–2929, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.
12. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, page arXiv:1810.04805, Oct 2018.
13. Huizhong Chen, Andrew C. Gallagher, and Bernd Girod. What's in a name? first names as facial attributes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.
14. H. Zhao and F. Kamareddine. 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 (CSCI)*, pages 222–227, Dec 2017.

15. Huiming Jin, Hao Zhu, Zhiyuan Liu, Ruobing Xie, Maosong Sun, Fen Lin, and Leyu Lin. Incorporating Chinese Characters of Words for Lexical Sememe Prediction. *arXiv e-prints*, page arXiv:1806.06349, Jun 2018.
16. Chuan Gu, Xi-ping Tian, and Jiang-de Yu. Automatic recognition of chinese personal name using conditional random fields and knowledge base. *Mathematical Problems in Engineering*, 2015, 2015.
17. Ming Liu, Vasile Rus, Qiang Liao, and Li Liu. Encoding and ranking similar chinese characters. *J. Inf. Sci. Eng.*, 33(5):1195–1211, 2017.
18. Gender Guesser, https://test.pypi.org/project/gender-guesser/. Last accessed 4 May 2019
19. Namsor Gender API, https://gender-api.com/. Last accessed 4 May 2019
20. Ngender, https://github.com/observerss/ngender/. Last accessed 4 May 2019
21. pypinyin, https://pypi.org/project/pypinyin/. Last accessed 4 May 2019
22. Gender Guesser, https://test.pypi.org/project/gender-guesser/. Last accessed 4 May 2019
23. Most common surnames revealed, http://www.chinadaily.com.cn/a/201901/31/WS5c528e7ea3106c65c34e78cb.html. Last accessed 4 May 2019