ISSN: 2641-2039

Global Journal of Engineering Sciences

DOI: 10.33552/GJES.2022.10.000740

Iris Publishers

Short Communication

A Novel Method On Translating People's Names In Mandarin - 'DT-NTM'

Hua Zhao* and Fairouz Kamareddine

School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK

***Corresponding author:** Hua Zhao, School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK, **Email :** dr.hua.zhao@gmail.com

Received Date: November 30, 2022

Published Date: December 07, 2022

Abstract

In-text translation, translating the romanised names to their original language names is challenging. Especially translating the accurate original Hanzi name from a Pinyin roman name. Moreover, our study found that most of the existing name translation methods only work with European languages. This paper focuses on Mandarin name translation. We propose a novel method that translates Pinyin names to Hanzi characters. We call it 'DT-NTM', which is a statistical method. The evaluation shows that 'DT-NTM' translates Mandarin names with a high degree of accuracy, 89.88%.

Keywords: Artificial intelligence; Nature language processing; Machine learning; Text translation

Introduction

In global communication and interaction, people's names are usually written in roman instead of their first language (e.g., Chinese, Japanese, Arabic, etc.). Here, we define 'Text diversity'[1] of people's names as the versions of a person's name in different characters. For example, '袁天罡' and 'Yuan Tian Gang' are a person's name but in different characters. Here, '袁天罡' is the person's name in Hanzi[2], and 'Yuan Tian Gang' is the Pinyin[3] version of '袁天罡'. Translating proper names[4] is a way to understand social events for social research.

However, translating proper names from one language to another is a challenge [1]. In this paper, we focus on the name translation in Mandarin[5] [2]. We propose a novel method to translate Pinyin names to Hanzi characters, 'DT-NTM' (Decision Tree - Name Translation Model). 'DT-NTM' is a statistical model. This paper's organisation is as follows: In section 2, we explain the proposed model 'DT-NTM'. Section 3 reports the evaluation results of 'DT-NTM'. Section 4 concludes our work in this article.

'DT-NTM' Overview

The process of 'DT-NTM'

There are two steps to translating a Pinyin name to Hanzi characters in 'DT-NTM'.

Step 1: We list all the possible Hanzi words of a proper Pinyin name from a Pinyin name N.

Step 2: We pick the suitable Hanzi character for the proper Pinyin name from a list of the corresponding Hanzi characters.

Methods in 'DT-NTM'

In this section, we will explain the methods in the second step of 'DT-NTM'. These methods assist in filtering the possible Hanzi words of a Pinyin name. We will split a Mandarin name into two parts to introduce these methods, 'filtering the Surname' and 'filtering the Given name'.

Methods of filtering a person's surname: Here are two methods that filter the possible Hanzi words of a Pinyin Surname for Mandarin name translation. These two methods are 'StartPro' and 'JudgProS'.

The methods of 'StartPro' and 'JudgProS' are shown as follows:

$$StartPro = \frac{C_i(s)}{\sum_{i=1}^{i} C_i(s)} \qquad (1)$$

$$JudgProS_{(s \vee g)} = \frac{C_i(s \vee g)}{C_i(s) + C_i(g)} \qquad (2)$$

Here, $C_i(s)$ is the frequency of the target Hanzi word 'C_i' as a Surname, which is from the designed training data set for 'DT-NTM'. And $C_i(g)$ is the frequency of the target Hanzi word 'C_i' as a Given name, which is from the

designed training data set for 'DT-NTM'.

Methods Of Filtering A Person's Given Name: The methods of 'JudgProF' and 'JudgProG' are used for narrowing the quantity of a list of possible Hanzi Given names $F_{(l,j)}$.

The methods of 'JudgProF' and 'JudgProG' are shown below:

$$JudgProF_{(s \vee g)} = \frac{F_{l,j}(s \vee g)}{F_{l,j}(s) + F_{l,j}(g)} \qquad (3)$$

$$JudgProG_{(m \vee f)} = \frac{F_{l,j}(gm \vee gf)}{F_{l,j}(gm) + F_{l,j}(gf)} \qquad (4)$$

Here, '$F_{l,j}$' is the frequency of the target $F_{l,j}(g)$ Hanzi word '$F_{l,j}$' as a Given name, which is $F_{l,j}$ from the designed training data set for 'DT-NTM'. And, '$F_{l,j}$' is the frequency of $F_{l,j}(s)$ the target Hanzi word '$F_{l,j}$' as a Surname, which is from the designed training data set for 'DT-NTM'. '$F_{l,j}$' is the frequency of the $F_{l,j}(gm)$ target Hanzi word '$F_{l,j}$' as a Male's Given name, which is from the designed training data set for 'DT-NTM'. '$F_{l,j}$' is the frequency $F_{l,j}$ (gf) of the target Hanzi word '$F_{l,j}$' as a Female's l,j Given name, which is from the designed training data set for 'DT-NTM'.

Result

In the evaluation of 'DT-NTM', we build three data sets. We use three open data sources to build these three data sets. The first data set6 covers 1,163,760 Han Chinese population. The second data set7 includes 1,163,760 names in Hanzi. The last corpus8 includes 4,900 Chinese male names and 4,900 Chinese females.

To test the proposed model of 'DT-NTM', we compare it with 'Google Translate'. In the evaluation, we use Accuracy (Acc) to report each model's performance using different lengths of people's names.

Table 1 indicates the evaluation results of 'DT-NTM' and 'Google Translate'. Based on this, the accuracy of 'DT-NTM' is 89.88%, and the accuracy of 'Google Translate' is 30.00%. In this experiment, 'Google Translate' is consistent and performs well when considering the average amount of the translated result of each proper Pinyin name.

Table 1: Comprehension Name Correct Response Rate of 'DT-NTM' and 'Google Translate'.

Data Set	DT-NTM (Acc (%))	Google Translate (Acc (%))
Two_word Name	63.39	14.74
Three_word Name	89.88	30.00
Four_word Name	50.00	10.71

Conclusion

This article has implemented a novel model for translating Mandarin names, 'DT-NTM'. We applied the testing data in the evaluation to test the proposed model and compare it with Google Translate. We have pointed out that this novel model has a good performance in people's name translation in Mandarin name, which has better accuracy than Google translate. In the future, we will study audio frequency for translating proper names.

Acknowledgement

None.

Conflict of Interest

No conflict of interest.

References

1. BS Pour (2009) How to translate personal names. Translation Journal 13(4): 1-13.
2. H Zhao, F Kamareddine (2019) A Novel Phonetic Algorithm for Predicting Chinese Names using Chinese Pin Yin. in MLDM, USA.

[1] Text diversity generally means that an idea has different world views expressions from different books and genres.

[2] Hanzi is a Chinese character.

[3] Pinyin is the translation into the Roman Alphabet of lexical tone transcriptions from Chinese characters.

[4] Proper name is a word which is the name of a person, a place, an institution, etc. It is written with a capital letter.

[5] An official language of Chinese which is spoken by over 730 million people.

[6] https://github.com/psychbruce/ChineseNames.

[7] https://github.com/wainshine/Chinese-Names-Corpus.

[8] https://www.researchgate.net/publication/2696305949800ChineseNameswithGender.