Chinese Pinyin Mapping Function

Contents

1. **Introduction**
2. **My Approach**
3. **Findings**
4. **Ideas for Further Research**
5. **Recommendations**

1. Introduction

The Chinese language is written using symbols called characters to represent meaning. Unlike words in a phonetic alphabet, there is no indication of how to pronounce each character. Therefore, to help people learn Chinese, a phonetic alphabet based on the latin alphabet was invented called pinyin. In different contexts, the same character can have different pronunciations, with some characters having as many as seven different pronunciations.

In textbooks, when a new character is presented, the pinyin is usually also given so that the reader learns how to pronounce the character. Chinese texts in the wild typically only contain characters. But for learners of the language, it is sometimes helpful to give the pronunciation of each character in a text. It is for this reason that I created a function that maps each character to it's correct pronunciation given the context.

2. My Approach

The first step in the process of building the pinyin mapping function was to generate a list of all the Chinese characters with all of their pronunciations and their frequencies. This required some parsing and pruning of a list of each character with its given pronunciation found at the Unihan Database. Unihan are Unicode for Chinese characters.

Some characters were missing frequency information so for these characters, I gave them a frequency of 1. Most of the characters that had multiple pronunciations had frequencies for them.

The second step was to build a mapping function that mapped single characters to their pronunciation based on the most frequent pronunciation of a given character. This approach would lead to correct mappings most of the time, but there are still some cases where the mapping would be wrong based on context. For this reason a more sophisticated mapping function was needed.

The final step in building a function that outputs the correct pinyin of each character was to use a phrase dictionary which contained the correct pronunciation of 96,809 words and phrases. I should note here that most Chinese words are made up of two to three characters. The majority of words and phrases in this list were two-letter words.

For the second function, I first looked at each set of two consecutive characters in the text and compared them to all of the words in the phrase dictionary to see if there was a match (ie. if these two characters were a word with a set pronunciation). If there was a match, the function output that pronunciation, if there was no match, then the pronunciation from the pinyin frequency dictionary was used. Then the function moves on to the next character in the text to repeat the process.

Sample input and output

Txt input:

中国历史上的重大事件，你知道几件？

如果直留下妹

发布时间：2021-12-02 16:51
关注
1、秦始皇统一中国

秦始皇是出生于赵国都城邯郸，前247年，13岁时即王位。22岁时，开始“亲理朝政”，自前230年至前221年，先后灭韩、赵、魏、楚、燕、齐六国，39岁时完成了统一中国大业，建立起一个以汉族为主体统一的中央集权的强大国家—秦朝，并奠定中国本土的疆域。秦王政灭六国后，认为自己“德兼三皇，功过五帝”，遂采用三皇之“皇”、五帝之“帝”构成“皇帝”的称号，是中国历史上第一个使用“皇帝”称号的君主，所以自称“始皇帝”。

2、安史之乱

安史之乱是中国唐代玄宗末年至代宗初年（755年12月16日至763年2月17日）由唐朝将领安禄山与史思明背叛唐朝后发动的战争，是同唐朝争夺统治权的内战，为唐由盛而衰的转折点。

这场内战使得唐朝人口大量丧失，国力锐减。因为发起反唐叛乱的指挥官以安禄山与史思明二人为主，因此事件被冠以安史之名。又由于其爆发于唐玄宗天宝年间，也称天宝之乱。

Txt output:

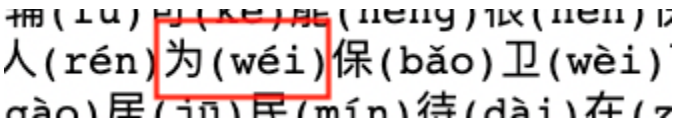
```
中(zhōng)国(guó)历(lì)史(shǐ)上(shàng)的(de)重(zhòng)大(dà)事(shì)件(jiàn)，你(nǐ)知(zhī)道(dào)几(jǐ)件(jiàn)？

如(rú)果(guǒ)直(zhí)留(liú)下(xià)妹(mèi)

发(fā)布(bù)时(shí)间(jiān): 2021-12-02 16:51
关(guān)注(zhù)
1、秦(qín)始(shǐ)皇(huáng)统(tǒng)一(yī)中(zhōng)国(guó)

秦(qín)始(shǐ)皇(huáng)是(shì)出(chū)生(shēng)于(yú)赵(zhào)国(guó)都(dū)城(chéng)邯(hán)郸(dān)，前(qián)247年(nián)，13岁(suì)时(shí)即(jí)王(wáng)位(wèi)。22岁(suì)时(shí)，开(kāi)始(shǐ)“亲(qīn)理(lǐ)朝(cháo)政(zhèng)”，自(zì)前(qián)230年(nián)至(zhì)前(qián)221年(nián)，先(xiān)后(hòu)灭(miè)韩(hán)、赵(zhào)、魏(wèi)、楚(chu)、燕(yàn)、齐(qí)六(liù)国(guó)，39岁(suì)时(shí)完(wán)成(chéng)了(le)统(tǒng)一(yī)中(zhōng)国(guó)大(dà)业(yè)，建(jiàn)立(lì)起(qǐ)一(yī)个(gè)以(yǐ)汉(hàn)族(zú)为(wéi)主(zhǔ)体(tǐ)统(tǒng)一(yī)的(de)中(zhōng)央(yāng)集(jí)权(quán)的(de)强(qiáng)大(dà)国(guó)家(jiā)—秦(qín)朝(cháo)，并(bìng)奠(diàn)定(dìng)中(zhōng)国(guó)本(běn)土(tǔ)的(de)疆(jiāng)域(yù)。秦(qín)王(wáng)政(zhèng)灭(miè)六(liù)国(guó)后(hòu)，认(rèn)为(wéi)自(zì)己(jǐ)“德(dé)兼(jiān)三(sān)皇(huáng)，功(gōng)过(guò)五(wǔ)帝(dì)”，遂(suì)采(cǎi)用(yòng)三(sān)皇(huáng)之(zhī)“皇(huáng)”、五(wǔ)帝(dì)之(zhī)“帝(dì)”构(gòu)成(chéng)“皇(huáng)帝(dì)”的(de)称(chēng)号(hào)，是(shì)中(zhōng)国(guó)历(lì)史(shǐ)上(shàng)第(dì)一(yī)个(gè)使(shǐ)用(yòng)“皇(huáng)帝(dì)”称(chēng)号(hào)的(de)君(jūn)主(zhǔ)，所(suǒ)以(yǐ)自(zì)称(chēng)“始(shǐ)皇(huáng)帝(dì)”。


2、安(ān)史(shǐ)之(zhī)乱(luàn)

安(ān)史(shǐ)之(zhī)乱(luàn)是(shì)中(zhōng)国(guó)唐(táng)代(dài)玄(xuán)宗(zōng)末(mò)年(nián)至(zhì)代(dài)宗(zōng)初(chū)年(nián)（755年(nián)12月(yuè)16日(rì)至(zhì)763年(nián)2月(yuè)17日(rì)）由(yóu)唐(táng)朝(cháo)将(jiàng)领(lǐng)安(ān)禄(lù)山(shān)与(yǔ)史(shǐ)思(sī)明(míng)背(bèi)叛(pàn)唐(táng)朝(cháo)后(hòu)发(fā)动(dòng)的(de)战(zhàn)争(zhēng)，是(shì)同(tóng)唐(táng)朝(cháo)争(zhēng)夺(duó)统(tǒng)治(zhì)权(quán)的(de)内(nèi)战(zhàn)，为(wèi)唐(táng)由(yóu)盛(shèng)而(ér)衰(shuāi)的(de)转(zhuǎn)折(zhé)点(diǎn)。

这(zhè)场(chǎng)内(nèi)战(zhàn)使(shǐ)得(de)唐(táng)朝(cháo)人(rén)口(kǒu)大(dà)量(liàng)丧(sàng)失(shī)，国(guó)力(lì)锐(ruì)减(jiǎn)。因(yīn)为(wèi)发(fā)起(qǐ)反(fǎn)唐(táng)叛(pàn)乱(luàn)的(de)指(zhǐ)挥(huī)官(guān)以(yǐ)安(ān)禄(lù)山(shān)与(yǔ)史(shǐ)思(sī)明(míng)二(èr)人(rén)为(wéi)主(zhǔ)，因(yīn)此(cǐ)事(shì)件(jiàn)被(bèi)冠(guàn)以(yǐ)安(ān)史(shǐ)之(zhī)名(míng)。又(yòu)由(yóu)于(yú)其(qí)爆(bào)发(fā)于(yú)唐(táng)玄(xuán)宗(zōng)天(tiān)宝(bǎo)年(nián)间(jiān)，也(yě)称(chēng)天(tiān)宝(bǎo)之(zhī)乱(luàn)。
```

3. Findings

In order to evaluate how my pinyin mapping function worked, I passed it three texts and checked them to see if there were any mispronunciations. Below are some of the pronunciations that were incorrect.

Article 1: New York Times Chinese Edition - Ukraine
Link: https://cn.nytimes.com/world/20220226/ukraine-russia-war-kyiv/

Errors:



Should be fourth tone: wèi

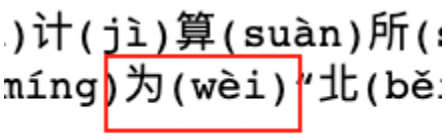
```
g)对(duì)乌(wū)克(kè)兰(lán)政(zhèng)府(fǔ)作(zuò)出(chū)嘲(cháo)讽(fěng)，称(chēng)其(qí)为(wèi)“一(yī)群(qún)瘾(yǐn)君(jūn)子(zǐ)和(hé)新(xīn)纳(nà)粹(cuì)分(fèn)子(zǐ)”。目(mù)前(qián)还(hái
```

Should be second tone: wéi

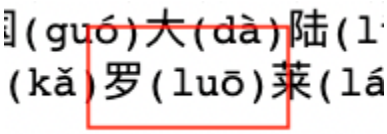
Total errors: 2

Article 2: Wikipedia - Lenovo
Link: https://zh.wikipedia.org/wiki/%E8%81%94%E6%83%B3%E9%9B%86%E5%9B%A2

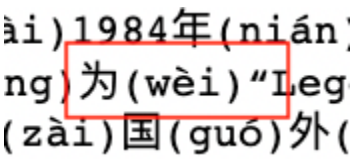
Errors:



Should be second tone: wéi



Should be second tone: lúo

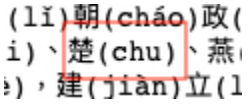


Should be second tone: wéi

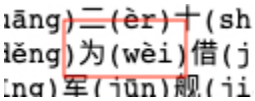
Total errors: 3

Article 3: Baidu - Chinese History
Link: https://baijiahao.baidu.com/s?id=1718023695121784419&wfr=spider&for=pc

Errors:



Should be third tone: chǔ



Should be second tone: wéi

Besides these few pronunciations, all of the other pronunciations were correct.

4. Ideas for Further Research

In order to build a more precise pinyin mapping function, the function could be made to check three-letter and four-letter words and phrases first using the phrase dictionary, and then proceeding to check for two-letter words and then single-characters words like in the function I created.

One could also present the pinyin and character of the output in a more attractive way. This project on Github presents some options for doing this.

My function only deals with simplified Chinese characters, which are used primarily in mainland China. But other places like Taiwan and Hong Kong use traditional Chinese characters, so a function could be built that would take these other forms of characters into account. Not all characters have a simplified version, by some estimates, 30% of the 3,500 most common characters in use have been simplified.

I could find a way to eliminate the incorrect pronunciations I found, especially wéi/wèi, which was responsible for over half of the errors I found.

5. Recommendations

The pinyin creator can be used to give the correct pronunciations to any simplified Chinese text passed to it. Three ways I recommend using this tool are:

1. It can be used to study the language by giving readers the pronunciations of unknown characters.
2. It can be integrated into an app or dictionary.
3. It can be used to cross check other pinyin mapping texts, apps, and dictionaries.