Pinyin to Hanzi Conversion with Hidden Markov Models

Lucas Freitas and Cynthia Meng

May 7, 2013

1. Introduction to Pinyin

The Chinese language is composed entirely of characters, with each character representing generally at least one word. When different characters are combined, different words might also be formed. Before modern technology, characters were written solely with brush strokes. During the 1950s, however, a system called *pinyin* was developed to aid the transcription of Chinese characters into the Latin alphabet.

Pinyin proved incredibly useful during the latter half of the century, as it was used to help teach the language to non-native Mandarin speakers, and eventually would be used as the main input method for entering Chinese characters into computers. It is still used today in Chinese language programs around the world, as it is one of the most efficient methods of teaching Chinese to native speakers of Latin-alphabet-based languages, such as English, Spanish, and others.

As a brief overview of Pinyin (of which a small amount of knowledge will help with the understanding of this project), there are five different tones in Mandarin Chinese, which can be represented by using the tone marks above vowels in the Pinyin of a character (Table 1).

Tone	Pinyin	Meaning
Flat or level (first)	mā	mother
Rising (second)	má	hemp
Falling-Rising (third)	mà	horse
Falling (fourth)	mă	scold
Neutral	ma	question

Table 1: The Mandarin Chinese four tones

2. Inspiration

Alhough Pinyin is certainly extremely helpful for visualizing the characters, some problems arise when attempting Pinyin to Hanzi conversion, especially given the breadth and scope of the Chinese language.

Both of us have a strong interest in the Chinese language and the relationship between characters and Pinyin, and were interested in implementing a project that would somehow draw these two integral components of the Chinese language together.

Inspired by the word segmentation problem set from earlier on in the course, we decided to attempt an implementation of Pinyin to Hanzi conversion, which involves both hempsegmentation and the use of a statistical model.

3. Heterophones

In Chinese, a given syllable could have up to five different tones (tones 1 to 4 or the neutral tone), each of which could stand for a different character. Especially when a piece of text is presented without tones over each word, it becomes exceedingly difficult to discern which character is the correct one to use. In this case, using context clues is particularly helpful, which our program would need to accommodate. Examples of this phenomenon

Pinyin (no tone)	Character	Meaning
hen	恨	to hate
hen	很	very
shi	十	ten
shi	是	to be
he	和	and
he	喝	to drink

Table 2: The Mandarin Chinese four tones

are in Table 2, with the pinyin words "hen", "shi", and "he".

0.1 Multiple Pronunciations for a Character

In addition to the problem of heterophones, there exists also the problem that certain characters can have different pronunciations depending on their contexts. We can see this with the following words:

wo jian ta le - 我见他了 (I saw him)
wo shou bu liao - 我受不了 (I can't stand it anymore)
yin hang - 银行 (bank)
zi xing che - 自行车 (bicycle)

0.2 Our Goals

Given these difficulties with Pinyin, we decided that our goal for this project would be to implement a system that, given a text composed of unsegmented pinyin, would return the proper Chinese character conversion of that Pinyin. For example,
Input:

mamamama, matimama

Output:

妈妈骂马, 马提妈妈
(Mother scolds the horse, the horse kicks Mother)

We would work with the Simplified version of Chinese, although we could likely easily adapt our program to account for Traditional form as well.

Possible Methods

Before beginning with our implementation, we brainstormed several different methods that could be used to implement this program. Among the factors we took into account were efficiency (both for the coders and the users), speed of performance, and compatibility with the problem we were trying to solve.

0.3 TANGO Algorithm

Obviously one of our first considerations was the TANGO algorithm that we had looked at in class and for Problem Set 3. In fact, the TANGO algorithm itself was actually developed for use with Japanese *kanji* segmentation. We considered perhaps that we could use this algorithm to help segment the Pinyin into the proper syllables, but ultimately decided that the algorithm fared better for simple segmentation rather than our goal, which was to both segment and probabilistically determine the correct character based on a given Pinyin.

0.4 Hidden Markov Model

The Hidden Markov Model was another consideration that we had, and it seemed to fit our goals quite well. The logic behind an HMM is that there are certain states in a process that are hidden to the the user(s); that is, we can see an output that could come from multiple different inputs, but we do not know the input from which the output is generated. What we do know, however, are the probabilities with which the possible input generates the output.

This obviously aligns quite well with our goals regarding the project. Given that a certain Pinyin could come from multiple different characters (generally, at least four different possibilities – at minimum!), we are trying to decipher which character is most likely to be the best fit based on the context of the sentence. The following diagram shows a rather simplified version of an HMM: Given the Pinyin "ni", we would first try and figure out which character is most likely to be the

correct one (there are actually far more than two possibilities, but we have only shown two here for simplicity's sake: 你, which means "you", and 泥, which means "dirt"). We would do the same for "hao", with the two characters here being 好, which means "good", and 号, which means "number" .

In the end, we would of course want the pairing 你好, which is the common way to say "hello" in Chinese.

0.5 Segmental Hidden Markov Model

Another option we considered, but ultimately rejected, was the Segmented Hidden Markov Model. The SHMM is very similar to the HMM, but

Implementation

Results

5.1 Test data

爱丽丝靠着姐姐坐在河岸边很久了由于没有什么事情可做她开始感到厌倦她一次又次地瞧瞧姐姐正在读的那本书可是书里没有图画也没有对话爱丽丝想要是一本书里没有图画和对话那还有什么意思呢

5.2 Output

爱丽丝靠着姐姐坐在河岸边很久了游于没游什么什情可坐她开什感到厌倦她易次游次第瞧瞧姐姐正在读的那本书可什书丽没游兔画也没游对画爱丽丝想么什易本书丽没游兔画河对画那还游什么易丝呢

5.3 Precision

0.712643678161

Analysis

Conclusion