Professor or screaming beast? Detecting Words Misuse in Chinese

Wei Liu, Ben Allison, Louise Guthrie

Department of Computer Science, University of Sheffield
w.liu@dcs.shef.ac.uk, b.allison@dcs.shef.ac.uk, l.guthrie@dcs.shef.ac.uk

Abstract

The Internet has become a very popular platform for communication around the world. However because most modern computer keyboards are Latin-based, Asian language speakers (such as Chinese) cannot input characters (Hanzi) directly with these keyboards. As a result, methods for representing Chinese characters using Latin alphabets were introduced. The most popular method among these is the Pinyin input system. Pinyin is also called "Romanised" Chinese in that it phonetically resembles a Chinese character. Due to the highly ambiguous mapping from Pinyin to Chinese characters, word misuses can occur using standard computer keyboard, and more commonly so in internet chat-rooms or instant messengers where the language used is less formal. In this paper we aim to develop a system that can automatically identify such anomalies, whether they are simple typos intentional substitutions. After identifying them, the system should suggest the correct word to be used.

1. Introduction

A certain kind of derogatory opinion is being conveyed in Chinese chat forums through the use of Chinese Hanzi (hieroglyphic) characters. There is potential for this to happen whenever two expressions are pronounced in a similar way in Chinese. Taking the example from the title, irate students have used "叫兽" ("Jiao Shou") for "教授" ("Jiao Shou"). While "教授" means professor, "叫兽" means "screaming beast" – clearly, we are in the presence of a disgruntled student!

A simple version of this phenomenon might occur in English as well, (say using Bezerkly for Berkley), but it is the method with which Chinese characters are input into a computer (by first typing the romanised version and then choosing the correct Hanzi) , which provides an easy opportunity for this kind of substitution to reflect feelings and prejudices of the writer.

Unlike the English example above, where a spelling detection program might be used to alert the user that a word substitution is being used, detection in Chinese is much more complicated, since there are no word boundaries. Segmentation programs will identify the two characters "教授" (meaning professor) as a word, and will correctly segment the phrase "叫兽" (screaming beast) into two words, so there is no immediate indication that something unusual appears in the text. As part of a project to detect anomalous words or phrases in Chinese, we have developed an initial algorithm for automatically detecting occurrences of this phenomenon. This paper describes the problem and evaluates the algorithm we have developed on a large body of chat forum data.

1.1. Background

There are 7 main groups of traditionally recognized dialects of spoken Chinese. The standardized form of spoken Chinese is Standard Mandarin. It is the official language of the People's Republic of China and the Republic of China (Taiwan), as well as one of the official languages of Singapore. In 1956, a Romanisation for standard Mandarin was introduced: it is known as "Hanyu Pinyin", or "Pinyin" for short. It is now the official phonetic transcription of mandarin Chinese.

In addition to Pinyin, there are several other Romanisations of Chinese, for example, the Wade-Giles[1], which was invented by Thomas Wade in 1859 and modified by Herbert Giles in 1892. Table 1 shows a comparison of the two popular romanisation schemes. For the purpose of this work, we use the Pinyin Romanisation: it is used by over 1.4 billion people and it is the standard phonetic representation of Chinese in mainland China.

1.2. Chinese Pinyin Input Systems

Chinese words are formed using Chinese characters (Hanzi). A Chinese word can range from one character long to four or more (if proper nouns are considered) characters long. To type Chinese into a computer, one needs to "translate" each character or word to a Romanised representation. The most popular method to do this is by using the Pinyin(Yuan, 1997). Every Chinese character has an associated Pinyin sequence: however, it should be remembered that as well as a single Pinyin corresponding to many characters, it is also true that a single character can have multiple Pinyins, depending upon context and thus semantics(Qiao et al., 1990)(Li and Grefenstette, 2005).

The simplest way of entering Pinyin and producing characters is sequential: a user types Pinyin syllables and selects the appropriate character after the completion of each syllable. For example, if a user wants to spell the Chinese word "China"(中国 Zhong Guo), he will type in the Pinyin of the first character "zhong",

and the input system will display a list of Chinese characters that all share that Pinyin (see figure 1). After he selects the correct character, he can then continue to type the Pinyin of the second character "guo" and again select the correct character.

Most modern Pinyin input methods attempt to match character sequences to Pinyin sequences (rather than treating each character individually, see figure 2), and because of the many choices corresponding to a Pinyin sequence, it

[1] http://www.pinyin.info/romanization/wadegiles/

Characters (Simplified/Traditional)	Hanyu Pinyin	Wade-Giles	Explanation
中国/中國	Zhōngguó	Chung1-kuo2	China
北京/北京	Běijīng	Pei3-ching1	Capital of the People's Republic of China
台北/臺北	Táiběi	T'ai2-pei3	Capital of the Republic of China in Taiwan
毛泽东/毛澤東	Máo Zédōng	Mao2 Tse2-tung1	Former Communist Chinese leader

Table 1: Mandarin Chinese Romanisation Schemes

Figure 1: Typical Chinese input system (by Google), character selection

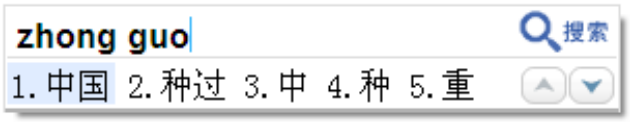

Figure 2: Typical Chinese input system (by Google), word selection

is easy for a user to make an incorrect choice(Lee et al., 1999), or to deliberately choose characters that sound close to the intended word, but have very different meaning. It is this kind of anomalous usage that we would like to detect automatically.

2. Word Anomaly in Chinese

Anomalies of the type described above are relatively easy to spot for native Mandarin Chinese speakers because both the intended word and substituted character sequence have very close or identical pronunciation.
In the following two examples, we illustrate the phenomenon:

1. 霉国(美国)要利用台湾问题来限制中国的快速发展。The U.S. wants to restrict China's fast development by taking advantage of the situation between mainland China and Taiwan.

2. 中国最恶心人的人群就是改革后的这些叫兽(教授)们。The most disgusting group of people in China are the professors after the reform.

These word anomalies are especially prevalent in on-line messages such as online discussion forums or instant messenger communications. In the first sentence, the word "美国"(The U.S.) is replaced with "霉国"; in the second sentence, The word "教授" (professor) is replaced by "叫兽". These are interesting in that although the replacement is not a single semantic unit, they are not typos but rather intentional substitutions. In 1 although "霉国" is not a semantic unit, if we consider the two characters separately, they are "霉"(fungus, moulded) and "国"(country). Thus the two-character sequence "霉国" means something like "fungus country" or "moulded country". Similarly, the character sequence "叫兽", interpreted one character at a time, gives the meaning "screaming beast".
These substitutions are intentional in that they express some sentiment (in this case, and most others, that sentiment is negative) towards someone or something. In 1 the speaker is blaming the U.S. government for what is considered to be misconduct towards the rather thorny mainland China/Taiwan situation. In 2 the speaker is accusing the professors who gained financially during China's economic reform of being profit-hungry, rather than displaying the virtues typically considered scholarly.

3. Related Work

Compared to automatic word error correction in English, research and work in Chinese automatic word error correction is very limited, possibly because the language presents added difficulties. One is due to a characteristic of the Chinese language which is that there are no explicit word boundaries. Although automatic word segmentation techniques are available, they are typically dependent on a lexicon. Moreover, the accuracy of automatic segmentation is not perfect, therefore, if the technique fails to recognise a word, the segmentation program will conclude that each single character is a word. Another difficulty occurs because the distinction between a non-word (such as "xyrt") and a context dependent error (such as using "there" for "their") that appears in English, is not present in Chinese. In contrast, almost every individual characters in Chinese is able to be a word on its own, and without taking context into account, all character
strings in any segmented Chinese text are all valid words. Therefore, if we adopt a dictionary approach that compares a list of all Chinese characters, then we will find virtually no "non-word" errors because each character may well be a word.
One representation method for automatic Chinese word correction was developed by(huang Chang, 1994) which uses confusing character substitution. In the method, confusing characters are used to replace every character in a given input sentence, and a "correct" result with highest evaluation score is searched from all paths. The obvious disadvantage of this method is that only errors caused by character substitution can be detected. Cai (Cai, 1997) found that most Chinese word errors cause segmentation

abnormality, for example segmentation of "忠耿耿", which is a character deletion error of the correct word "忠心耿耿" (be devoted), is likely to be three single-character words "忠/耿/耿". Native speakers can identify that "忠耿耿" is a non-word with little effort, but it is not possible for a computer to determine that it is a non-word error (rather than three single-character words).

Zhang et al. (Zhang et al., 2000a) proposed an approximate word matching algorithm and complex word substitution method to detect and correct Chinese word errors. They suggest that Chinese text errors should be categorised as non-word errors and real-word errors similar to English. Non-word errors refer to a string that cannot be treated as a word in the error text, and these errors include character substitution errors, such as "厉害"→"历害", "一鸣惊人" → "一呜惊人", character insertion error such as "惊天动地"→"惊天天动地" and character deletion errors "忠心耿耿"→"忠耿耿". Real-word error, similar to English, are those error strings resulting in another valid word(s); they can be character substitution errors, such as "通知"(notice) →"同志"(comrade), or insertion errors, such as "基于"→"基于基于", or character deletion errors: "仙人掌"(cactus)→"仙人"(immortal).

The paper also stated that "approximate word match method is used in English spelling error detection and correction to find all words in a dictionary whose minimum edit distance to a given string is less than a threshold"(pp. 249). They also acknowledged that approximate word match of Chinese is considerably different; accordingly they proposed an approximate Chinese word match method to find all words that have a distance less than a threshold to the sub-strings. The method begins from a specific position in the target sentence, and find all the sub-strings with varying length. the position with different length 1,2,3,... from a specific position in the target sentence.

Another attempt by (Zhang et al., 2000b) claimed that most Chinese word correction techniques performed poorly because they adopted a rather naïve language model that only considered character or word n-grams, they claimed that these models can only present local language constraints. As a result they presented a Winnow-based error detection and correction approach that used both local language features as well as wide scope semantic features. The features included: words before and after the target strings, part-of-speech trigram, context word semantic category according to 《同义词词林》(Synonyms dictionary), and characters within words. They trained on news texts consisting of 100 million characters and tested with 335 real word errors caused by the "five-stroke" input. The method managed to achieve 88% on recall, 64% on precision in error detection and 56% with the correction rate, compared to the trigram model which obtained 69% on recall and 35% on precision.

4. Automatically Detecting Anomalies

We have developed a method to detect phonetic anomalies that have the following properties:

1. The anomalous character sequence and the intended one both have similar pronunciations. To start with, we consider their pronunciation to be similar if they both have identical Pinyin (where we use the more standard form of Pinyin without tonal information).
2. The anomalous character sequence and its original are at least 2 characters long
3. The anomalous character sequence is not listed as a single entry in a standard lexicon, so its only meaning is combinatoric.

We divided the identification process into three steps as below:

- Step 1 - Segmentation

Given a piece of Chinese text, we first feed it into an automatic word segmenter(Zhang et al., 2003) to break the text into semantic units. Because we consider only multiple-character anomaly cases, anomalies can only be contained within sequences of single characters (the segmenter uses a lexicon, and thus will only segment into words sequences which appear within that lexicon).

- Step 2 - Character sequence extraction

After segmentation, we are interested in sequences of single characters, because anomalies will occur only within those sequences. Once we obtain these sequences, we generate all possible sub-strings for each sequence because any anomalous words can be part of a character sequence.

- Step 3 - Detection

We assume the anomaly shares many phonetic similarities with the "true" word. As a result we need a method for comparing pronunciations of two character sequences. Here we use the Pinyin to represent phonetics of a Chinese character, and in this early stage we define two pronunciations to be similar when they both have identical Pinyin (not including the tone).

We obtain a freely available character-to-pinyin conversion tool[2] that can produce all possible Pinyin sequences corresponding to a given Chinese character. Using this tool we create a Pinyin-to-Word hash table using the machine-segmented Chinese Gigaword version 2. The keys of the hash table are Pinyin sequences and the values are words with that Pinyin. With this Pinyin-to-Word hash table we can fast search for a word given its Pinyin.

Once we have the resources, we first produce all possible Pinyin sequences of each character sequence (step 1). We do this for each of the sub-character sequences extracted by step 2. After obtaining the Pinyin, we then do a Pinyin-word look up in the hash table we created; if there exists any entries, we know that the Pinyin sequence maps to one or more real words. These mappings could possibly be the original, intended words. Consequently, we consider any character sequences whose Pinyin maps to real words to be possible anomalies.

At this preliminary stage, if there are multiple candidates, we will consider the most frequent word in the Gigaword corpus to be the true word.

[2] http://pinyin4j.sourceforge.net/

Correct/Intended word	Misused character sequence	Pinyin	Occurrence
美国 (The U.S.)	霉国 (fungus country)	Mei guo	43
教授 (Professor)	叫兽 (screaming beast)	Jiao shou	23
偶像 (Role model)	呕像 or 呕象 (people that make one sick)	Ou xiang	12

Table 2: Testing document

- Step 4 - Further tuning

Because many words in Chinese are single character words, there can be many single-character word sequences that our method will capture. Furthermore, because of the highly ambiguous Pinyin-to-character mapping, step 3 is likely to over-generate candidates which are not misused character sequences, but whose Pinyin sequences happen to be identical to alternative character sequences in the lexicon. To reduce these falsely identified character sequences (false positives), we adopt the following strategies.

1. We assume each character sequence contains at most one misused word; if many are identified the longest character sequence is preferred. Consider that we encounter a character sequence "霉国人", which is a misused character sequence for the word "美国人"(American), because both "美国"(the U.S.) and "美国人"(American) are lexicon listed words, our identification process stated in step 3 will decide that the two character sequences, "霉国人" and "霉国" are candidates of misused. In this case, we only select the longer candidate "霉国人" to be the misused character sequence.

2. A record of previously identified misused character sequences is kept. If multiple candidates for misused character sequences exist, we will prefer the one that had been previously identified, i.e. in the cache.

3. We assume that words which are replaced to form misuses are commonly used words. We consider a character sequence to be misused only when its alternative-word bigram appeared above a pre-set threshold in a large corpus. Thus we built a bigram word-frequency count using the whole Chinese Gigaword version 2. When we identified misused character sequence candidates in step 3, we consider it as a possible alternative word. We then form a bigram from this word and the one immediately following, as shown in figure 3. We look this bigram up in our indexed Gigaword corpus to see how many times this bigram has occurred. Because we assume the intended word will be fairly common, we set a threshold and only consider the character sequence to be misused if the bigram occurrence is above the threshold. In our experiments we will vary this bigram threshold to attempt to improve performance.

5. Data and Experiments

We have conducted preliminary experiments to test our algorithm. To start with, we manually gathered a small number of documents which contain anomalous phrases of the type described above. The documents are gathered from internet chat-rooms and contain 3,797 Chinese characters: the anomalies herein are shown in table 2.

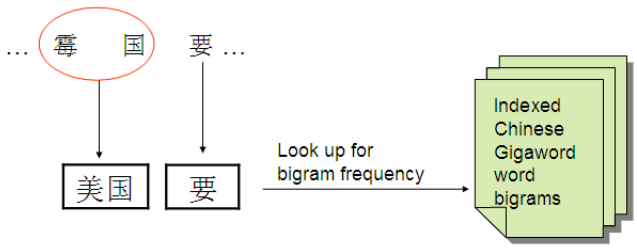

Figure 3: Bigram look up

5.1. Results and Discussion

We evaluate our identification/correction performance using standard measures of precision and recall, let a be the number of misused words correctly identified and corrected by our method; let b be the total number of character sequences our method returned as misused words; and c be the totally number of misused words in the test data. so we can defined precision (p) and recall (r) as (1) and (2). We tested our performance using 1, 2 and 3 bigram thresholds.

$$p = a/b \quad (1)$$

$$r = a/c \quad (2)$$

Table 3 shows the performances of our method. Because the size of the testing data is small, we manually checked the output produced by our method in order to determine correctly identified and corrected anomalous words.

	Bigram Threshold		
	≥1	≥2	≥3
No. of misused chararcter sequence	78	78	78
System identified	130	93	61
Correctly identified	78	69	48
Precision	60%	74.19%	78.69%
Recall	100%	88.46%	61.54%
F-measure	75%	80.7%	69.07%

Table 3: Result for word misused identification

The initial experiments showed that our method can successfully identify and correct the three examples of non-word anomalies with reasonable precision and recall. The basic method obtains 100% recall however it generates a lot of false positives; this can be seen in a relatively low precision of 60%.

In summary, our method is successful at identifying genuine anomalous non-word character sequences; however

the method also retrieves some false positives, due to the highly ambiguous Pinyin to word mappings.

6. References

Sun Cai. 1997. Research on lexical error detection and correction of chinese text. Master's thesis, Tsinghua University, Beijing, China.

Chao huang Chang. 1994. A pilot study on automatic chinese spelling error correction. *Communication of CPLIPS*, 4(2):143–149.

Kin Hong Lee, Mau Kit Michael Ng, and Qin Lu. 1999. Text segmentation for chinese spell checking. *Journal of the American Society for Information Science*, 50(9):751–759.

Yiping Li and Gregory Grefenstette. 2005. Translating chinese romanized name into chinese idiographic characters via corpus and web validation. In *Proceedings of CORIA 2005*, Grenoble, France.

Jinan Qiao, Yizheng Qiao, and Sanzheng Qiao. 1990. Six-digit coding method. *Commun. ACM*, 33(5):491–494.

Chen Yuan. 1997. *Chinese Language Processing*. Shang Hai education publishing company.

Lei Zhang, Changning Huang, Ming Zhou, and Haihua Pan. 2000a. Automatic detecting/correcting errors in chinese text by an approximate word-matching algorithm. In *ACL'00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 248–254, Morristown, NJ, USA. Association for Computational Linguistics.

Lei Zhang, Ming Zhou, and Changning Huang. 2000b. Approach in automatic detection and correction of errors in chinese text based on feature and learning. In *The Third Chinese World Congress on Intelligent Control and Intelligent Automation (In Chinese)*, pages 2744–2748.

Hua-Ping Zhang, Qun Liu, Xue-Qi Cheng, Hao Zhang, and Hong-Kui Yu. 2003. Chinese lexical analysis using hierarchical hidden markov model. In *Proceedings of the second SIGHAN workshop on Chinese language processing*, pages 63–70, Morristown, NJ, USA. Association for Computational Linguistics.