Detecting Word Misuse in Chinese

Wei Liu
Department of Computer Science
University of Sheffield
W.Liu@dcs.shef.ac.uk

Abstract

Social Network Service (SNS) and personal blogs have become the most popular platform for online communication and sharing information. However because most modern computer keyboards are Latin-based, Asian language speakers (such as Chinese) has to rely on a input system which accepts Romanisation of the characters and convert them into characters or words in that language. In Chinese this form of Romanisation (usually called Pinyin) is highly ambiguous, word misuses often occur because the user choose a wrong candidate or deliverately substitute the word with another character string that has the identical Romanisation to convey certain semantics, or to achieve a sarcasm effect. In this paper we aim to develop a system that can automatically identify such word misuse, and suggest the correct word to be used.

1 Introduction

A certain kind of derogatory opinion is being conveyed in Chinese chat forums and SNS sites 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. For exmaple, irate readers have used “妓者” (”Ji Zhe”) for “记者” (”Ji Zhe”). While “记者” means reporter or journalist, “妓者” can be interpreted as prostitute.

There are 5000 commonly used characters. While the number of distinct Pinyin (toneless) is only 412. Therefore Pinyin to character conversion is highly ambigurous and is a active research topic (Zhou et al., 2007), (Lin and Zhang, 2008), (Chen and Lee, 2000). On the other hand, automatic Pinyin generation is considered a solved task, (Liu and Guthrie, 2009) shows that using the most frequent Pinyin approach to assign Pinyin to each character can achieve 98% accuracy. In fact, we test on the Gigaword Chinese (Verson 2) corpus and find out that only about 15% of the characters have ambigurous Pinyin.

2 Automatically Detecting Word Misuse

We divided the detection process into three steps as below:

- 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.
- 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 substrings for each sequence because any anomalous words can be part of a character sequence.
- 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 we define two pronunciations to be similar when they both have identical Pinyin (not including the tone). We use character-to-pinyin conversion tool[1] to create a Pinyin-to-Word hash table using the machine-segmented Chinese Gigaword

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

Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, pages 5–6,
Los Angeles, California, June 2010. ©2010 Association for Computational Linguistics

ver. 2. Once we have the resources, we first produce all possible Pinyin sequences of each character sequence.Next we 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. Consequently, we consider any character sequences whose Pinyin maps to these words to be possible anomalies.

3 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 1.

Intended word	Misused character seq.	Pinyin	Freq
美国 (The U.S.)	霉国	Mei guo	43
教授 (Professor)	叫兽	Jiao shou	23
偶像 (Role model)	呕像or 呕象	Ou xiang	12

Table 1: Testing document

3.1 Results and Discussions

We evaluate our identification/correction performance using standard measures of standard precision and recall. We tested our performance using bigram thresholds of 0, 1 and 2.

Table 2 shows the performances of our method.

No. of misused chararcter sequence	78
Total identified	130
Correctly identified	78
Precision	60%
Recall	100%
F-measure	75%

Table 2: Result for word misuse 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 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.

4 Future Work

Our experiments shows that our preliminary method can detect word misuses due to the Pinyin sequence being idential but with a relatively high false positives. In the future we plan to use other contextual evidence, such as pointwise mutual information to model whether the candidate sequence generated by our method is a better fit than the original sequence. We also plan to gather more real data that contain misuse of our interests.

References

Chen, Z. and Lee, K.-F. (2000). A new statistical approach to chinese pinyin input. In *In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 241–247, Hong Kong.

Lin, B. and Zhang, J. (2008). A novel statistical chinese language model and its application in pinyin-to-character conversion. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1433–1434, New York, NY, USA. ACM.

Liu, W. and Guthrie, L. (2009). Chinese pinyin-text conversion on segmented text. In *TSD '09: Proceedings of the 12th International Conference on Text, Speech and Dialogue*, pages 116–123, Berlin, Heidelberg. Springer-Verlag.

Zhang, H.-P., Liu, Q., Cheng, X.-Q., Zhang, H., and Yu, H.-K. (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.

Zhou, X., Hu, X., Zhang, X., and Shen, X. (2007). A segment-based hidden markov model for real-setting pinyin-to-chinese conversion. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 1027–1030, New York, NY, USA. ACM.