Statistical Analysis on Large Scale Chinese Short Message Corpus and Automatic Short Message Error Correction

Rile Hu[1], Yuezhong Tang[1], Chen Li[1,2] and Xia Wang[1]

[1]NOKIA Research Center, Beijing
[2]Beijing University of Posts and Telecommunications

{rile.hu, yuezhong.tang, ext-chen.9.li, xia.s.wang}@nokia.com

Abstract. Analysis of short message corpus is an important foundation for research of automatic short message processing technology. Based on large scale short message corpus, this paper firstly presents statistical data and performs analysis in detail on basic information of short message corpus and special language phenomena in it. The distributions of the corpus parameters and special language phenomena are also given out. The statistical results presented in the paper are meaningful for research of robust short message understanding and implementation of short message based man-machine dialog system and short message based machine translation system. And we also build an automatic error correction system on mobile phone to correct the misapplication of Chinese character in short messages. The preliminary results show that our method is effective.

Keywords: computer application; Chinese information processing; corpus technology; statistical analysis; short message; error correction.

1. Introduction

With the development and popularization of mobile phones, the short message is widely used for communication and plays more and more important role in people's daily life. As a main part of mobile communication, the short message has become a hot spot of mobile communication service in China at present because of its huge quantity and rapid progress. According to the statistics, there are 0.4 billion mobiles and 300 billion short messages in 2005 in China. The short message is so widely used and rapidly communicated that it makes our lives much convenient. However, on the other hand, there are a series of problems on the use and application of the short message.

The limitation of mobile input leads to wrong words, self-made symbols, non-punctuation and many other non-standard Chinese problems in short message text at present. These non-standard words and application customs spread to all ages and many places with the fast and wide radiation of short messages, which makes a huge negative effect to correctly use and inherit Chinese. (For example, the students form all over the country use many short message words to write compositions of the entrance examination to college.)

With the development of Corpus Linguistics (Kennedy, 2000), more and more corpora are built for research works. The results of the research works are widely used in natural language processing tasks. For example, Brown University in US built BROWN corpus, Lancaster University in UK and Oslo University and Bergen University in Norway built LOB corpus (Johansson, 1991). In China, large scale corpora were built by Tsinghua University, Peking University, Chinese Academy of Sciences and Chinese Academy of Social Sciences. Statistic and analysis had been done on these corpora (Feng,1999) (Feng ,2002) (Yu, 1998) (Yu,2002) (Guo, 2005) (Hu, 2002). The works listed above are focused on written languages. There are also some research works on Chinese spoken language. Qualitative analysis had been taken on

22nd Pacific Asia Conference on Language, Information and Computation, pages 397–403

Chinese spoken language in early research (Chen, 1984). A spoken language corpus on specific domain was also built; statistic analysis had been taken on this corpus (Zong, 1999). But for the specific domain of short message, there is not any corpus built and analyzed.

The short message is characterized by informal, highly interactive and other features of oral Chinese. However, it has many differences with the normal oral Chinese. Therefore, how to establish an effective corpus analysis measures on exact statistics analysis is so important to understand and handle the short messages.

2. Basic Information of the Corpus

The corpus contains 410K short messages. And these messages contain more than 7M Chinese character. The average length of these short messages is 16.98 Chinese characters.

Automatic word segmentation and POS tagging tools (Zhang, 2007) are used here to process the short message corpus. According to the characteristic of the Chinese short message, we classify all the word in the corpus into 18 part-of-speeches: adjective (A), conjunction (C), adverb (D), position word (F), idiom (I), abbreviated word (J), common-used phrase (L), measurement word (M), noun (N), onomatopoeia (O), preposition (P), quantifier (Q), pronoun (R), location word (S), time word (T), auxiliary (U), verb (V) and particle (Y).

Statistical results are got from the corpus after segmentation and POS tagging. Some of these results are shown in the coming sub-sections.

2.1.Distributions of Word Length

The distributions of word length of the corpus are shown in Table1:

Table 1: the distributions of word length

Word length (characters)	1	2	3	More than 4
Percentage（%）	67.01	30.64	1.88	0.47

There are 1.31 Chinese characters per word in the short message corpus. The word length of short message is short than that of the spoken language (about 1.87 characters per word), and much shorter than that of the written language (about 2.45 characters per word) (Chen, 1984) (Zong et.al, 1999).

2.2.Distributions of Message Length

The distributions of message length of the corpus are shown in Table2:

Table 2: the distributions of message length

Length（characters）	1～10	11～20	21～30	31～40	41～50	51～60	61～70
Percentage（%）	31.27	46.68	12.81	4.54	2.25	1.28	1.17

The messages no more than 30 characters cover the more than ninety percent of the corpus. This shows that people lean to use short and simple message in mobile communication.

2.3.Distributions of Part-of-Speech

The distributions of part-of-speech of the corpus are shown in Figure 1:

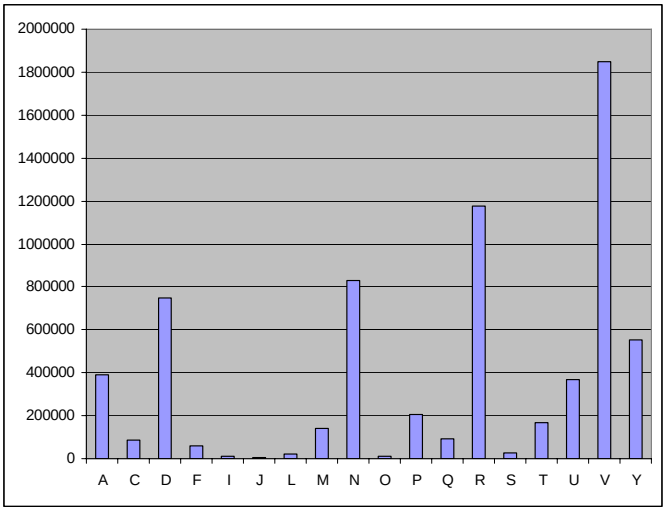

Figure 1: the distributions of part-of-speech

The highest frequent part-of-speeches are verb, pronoun, noun, adverb, particle and adjective. This is similar to spoken language (Zong et.al, 1999).

3. Special Language Phenomena Analysis

There are many special language phenomena contained in short messages, this makes the processing of short messages more difficult. The analysis of these phenomena is important for the development of short message processing technology.

Some manual labeling and modification of these special phenomena are made on the corpus. And also some statistical analysis are made on the manually processed results.

3.1.Classification of the Special Language Phenomena

The special language phenomena of the short message can be classified into the categories below:

（1）Usage of dialect words

Short message and spoken language have many similar characteristics, including the usage of dialect words. Some short messages contain some words and expressions in dialects; this brings difficulty for the automatic processing of short messages.

Example:

阿拉是上海人

Here, the word "阿拉" is a dialect word which is not used in Mandarin Chinese.

（2）Misuse of words with same pronunciation

Example:

你自几对我说　　（你自己对我说）

The character "几" and "己" have the same pronunciation, but their meanings are different. When users input short messages using pinyin input method, if they directly select the characters ranked high but not the actual character, they will misuse the words with same pronunciation. This phenomena also brings some difficulties for automatic short message processing.

（3）Misuse of words with similar pronunciation

Example:

我从启下机子。（ 我重启下机子 。）

Here, the character "从" (cong) and character "重" (chong) have similar pronunciation and the first one are misused instead of the second. This is because that some users in China could not distinguish the Pinyin（z、c、s）and（zh、ch、sh）, the Pinyin（in）and（ing）, and

etc. , so they uses the characters with similar pronunciation instead of the original ones. This also makes the processing of short message harder.

（4）Misuse of words with similar shape

Examples:

a) 这儿件事 (这几件事)

b) 自己 (自己)

This mistake occurs when the users choose the wrongly recognized character when using pen input method to input short messages. This will make the error correction and message understanding very difficult.

（5）Redundancy

Redundancy means that the short message contains redundant words.

Example:

你那儿里是不是有事？ (你那里是不是有事？)

（6）Incomplete word

This means that there are some incomplete words in short messages.

Example:

我去街买点东西 (我去街上买点东西)

（7）Repeat:

This means that short messages contain some similar meaning parts.

Example:

越来越愈来愈幽默了 (越来越幽默了)

（8）Usage of net languages

This means that short messages often contain some strings (maybe not real words) widely used in internet.

Example:

偶不知道 (我不知道)

This is because that these strings are widely used in some net forums and bbs, and the users take these strings into their daily life. So these strings are used and spread in short messages.

（9）Disorder

This means that there are some disorders in short messages.

Example:

为什么说五周见？ （为什么说周五见？）

Some of these special language phenomena are similar to spoken language (Redundancy, Repeat, Disorder and etc.), some of them are special in short message corpus (Misuse of words with same pronunciation, Misuse of words with similar pronunciation and etc.). These phenomena affect the automatic processing and understanding of the short messages very much. How to solve the problem is an important task.

3.2.Distributions of the Special Phenomena

The distributions of the special phenomena are shown in Table 3:

Table 3: the distributions of the special phenomena.

Phenomena No.	1	2	3	4	5	6	7	8	9
Percentage (%)	3.66	24.78	9.29	0.71	33.30	24.00	1.99	1.32	0.95

The results shows that the highest frequent special phenomena are: redundancy, misuse of words with same pronunciation and incomplete word.

Some of these phenomena (Misuse of words with same pronunciation, Misuse of words with similar pronunciation and Misuse of words with similar shape) are correlative to the input

methods of mobile phones. If some improvements are made to solve these problems in input methods, these errors could be perished.

4. Automatic Short Message Error Correction

We focus on the problem of the misuse of words with same pronunciation and the misuse of words with similar pronunciation; build an automatic error correction system on mobile phone. The system uses language model to correct the misapplication of Chinese character. And the preliminary results show that our method is effective.

There are three assumptions in advance:

1) The error in the Chinese character string comes not from the wrong input (For example, if "liu" should be input in the pinyin input approach, it is not "li", "lin" or some others but "liu" input) but from the wrong choice.

2) Most of the wrong Chinese characters cannot compose a word with the neighbour characters.

3) Most of the characters in the string are correct.

These 3 assumptions are usually assured in practice.

In the following, we will take Pin Ying input method as an instance. There should be a mapping table between Pin Yin string (PYMT), never wrong characters set (NWCS) and Chinese words and a language model (LM) based on Chinese words in advance. NWCS includes some characters which never be wrongly used in practical corpus. The algorithm is the same for all the input approaches. The processing procedure is listed as follows:

(1) Word segmentation

Firstly the sentence is segmented to some words group according to a pre-defined lexicon, which can be updated dynamically. According to the assumption 2, the word, which consists of 2 or more Chinese characters, should be correct in most situations, and the word that is a single Chinese character may be the wrong character. They will be treated in different way. The word, which consists of 2 or more Chinese characters, should be checked by language model. Those that have low score in language model will be labelled as doubtful. The word that is single Chinese character will be labelled doubtful except those that belong to NWCS.

(2) Recover to Pin Yin string

The doubtful Chinese words will be transformed to Pin Yin code. If the Chinese Character has more pronunciations, all of the possible Pin Yin codes should be kept. Moreover, the tone of Pin Yin will be removed because the user usually ignores tone while input characters. (The process of tone is particularly for Pin Yin input. For the other input method, it is ignored)

(3) Search Chinese Character based on the Pin Yin

The sentence with Pin Yin will be matched from the beginning to the end according to the PYMT. Any matched Chinese word from the table will be selected. By the way, PYMT can be extended so that some similar codes will be included. For example, "jin" and "jing" can be treated equally. With the existed Chinese words, they will compose several possible sentences. Now the possibilities of all the possible sentences will be calculated based on the language model. The top N best results will be kept. By the way, in calculating with language model, there should be threshold. The candidates which possibility is lower than threshold value should be given up. If all the possibility is lower than threshold, the original characters will be kept.

The language model and mapping table can be different in terminal and server.

(4) Select the best candidate

In the terminal of mobile phone, the users can help to select one from the Top N candidates.

The selection can be done automatically, especially in server. The selection criterion should consider both the possibility based on language model and the match degree with the original sentence, which is related with the assumption 3.

Here is a sample:

Original: 这次华费乐十元

Segment and Pin Yin transform: 这次 hua fei le 十元

Table match: 这次 花 费 了 十元，这次 话 费 了 十元，这次 花 费 乐 十元，这次 华 费 乐 十元 …

Language model check: 这次 花 费 了 十元 has the highest score. It will be chosen

Final result:这次花费了十元, this is what we want.

Now, there are 7 candidates of Chinese characters after inputting any code in mobile phone. If the correct Chinese character is not in the 7 candidates, you have to page down to search. Moreover, with the help of this technology, the input method can be adjusted. For any code, there are only 6 candidates of Chinese characters. The 7th is special character. If it is chosen, it means that the user prefers this character shown in the form of code (PY, WB, or stroke). The correct Chinese character will be found until the text modification is activated. In this way, the user reduces the burden of selection. Moreover the location of character that needs to be modified is clear. This is helpful to improve the modification accuracy.

The preliminary experiment has been done on the mobile phone platform. The accuracy of the system is 54.68%, the recall is 71.17%, and F measure is 70.86%. This is good since that the system runs on mobile phone, the resources are limited.

5. Conclusion

On the basis of the above-mentioned statistics and example analysis, we make the conclusions as follows:

1) The corpus of short message is similar with that of Chinese spoken language. Both of the words used are shorter than that of the formal language, among which the corpus of short message word is much shorter. The distributions of the part-of-speeches are similar.

2) Short message is highly interactive, which the pronoun and particle are frequently used.

3) Short message has many special language phenomena that make great effects on its automatic processing and understanding. The study and process of the corpus according to the special phenomena will benefit the short message use and spread.

We made a statistics analysis on large scale short message corpus in this paper, reflecting the rule of choosing words and making sentences in daily life, and all kinds of special language phenomena. These results are valuable to establish robust short message understanding system, short message based human-computer dialog system and machine translation system as reference.

And we also built an error correction system on mobile phone for Chinese short messages. The preliminary results show that our method is effective.

References

Chengqing ZONG, Hua WU, Taiyi HUANG and et al, 1999. Analysis of Spoken Dialog Corpus in Restricted Domain, *In Proceedings of the 5th Joint Symposium of Computational Linguistics.* Beijing: Tsinghua University Press, pp. 115-122.

George GUO, Hua LIU, Xuemin XIE and et al, 2005. The Initial Statistic Analysis on Tagged Corpus of People's Daily. *Natural Language Understanding and Large-scale Content Computing.* Beijing: Tsinghua University Press, pp. 187-192.

Graeme Kennedy, 2000. An Introduction to Corpus Linguistics. *Beijing: Foreign Language Teaching and Research Press.*

Guohua ZHANG. 2007. Research and Application of Automatic Chinese Word Segmentation and POS Tagging Algorithm. *Beijing: Graduate School of Chinese Academy of Sciences.*

Jianmin CHEN, 1984.Chinese Spoken Language. *Beijing: Beijing Press.*

Johansson S., Stenstorm, 1991. English Computer Corpora. *Berlin: Mouton de Gruyter.*

Rile HU, Chengqing ZONG, Juha Iso-Sipilä, et al, 2002.Investigation and Analysis on Designing Chinese Balance Corpus . *In Proceedings of the International Symposium on Chinese Spoken Language Processing (ISCSLP2002),* Taiwan：, pp. 335-338.

Shiwen YU, Xuefeng ZHU, Hui WANG and et al, 1998. Dictionary of Modern Chinese Grammatical Information. *Beijing: Tsinghua University Press*.

Shiwen YU, Huiming DUAN, Xuefeng ZHU and et al, 2002. The Basic Processing of Contemporary Chinese Corpus at Peking University Specification. *Journal of Chinese Information Processing*, Vol. 16, No. 5, pp. 49-64.

Yuan LIU, Qiang TAN, Xukun SHEN, 1994. The Standard of Chinese Word Segmentation for Information Processing and Automatic Word Segementation Methods. *Beijing: Tsinghua University Press.*

Zhiwei FENG, 2002. Evolution and Present Situation of Corpus Research in China. *Journal of Chinese Language and Computing*, Vol. 12, No. 1, pp. 43-62.

Zhiwei FENG, 1999. Corpus Linguistics and Machine Translation. *In Age of Information Network and Japan Research*, Shandong：Shandong University Press.