Character usage in Chinese Short Message Service (SMS): a real-world study in Mainland China

Xi Chen

School of Management,
Department of Management Science and Engineering,
Zhejiang University,
Administration Tower 1001,
Zi Jin Gang Campus, Hangzhou, China
E-mail: chen_xi@zju.edu.cn

Chenhui Guo

Department of Management Information Systems,
The University of Arizona,
McClelland Hall 430, 1130 E. Helen Street,
Tucson, AZ 85721, USA
E-mail: chguo@email.arizona.edu

Michael Chau

School of Business,
Faculty of Business and Economics,
The University of Hong Kong,
Pokfulam Road, Hong Kong
E-mail: mchau@business.hku.hk

Weihua Zhou*

School of Management,
Department of Management Science and Engineering,
Zhejiang University,
Administration Tower 1001,
Zi Jin Gang Campus, Hangzhou, China
E-mail: larryzhou@gmail.com
*Corresponding author

Abstract: Short Message Service (SMS) is an important component of modern mobile services. Given unique characteristics of Chinese language, it is imperative to conduct study to understand characteristic of language usage patterns in Chinese SMS so that important facts like why and how people in China use SMS can be discovered. In this paper, we report an analysis of Chinese SMS logs from three different provinces in China. A computational approach was applied to extract *n*-grams from logs of SMS. The language usage patterns reported in this paper consist of two aspects: 1) most popular *n*-grams that represent what types of information were transmitted via SMS;

2) distribution of *n*-grams in comparison with Zipf laws. We discovered that, compared with other forms of free text in Chinese, SMS contains more conversational elements, which are expressed mostly in bigrams. Trigrams, 4- and 5-grams are less frequent but are closely connected to commercial activities, which may indicate the commercial needs of SMS users.

Keywords: mobile communication; SMSs; short message services; character usage; mobile services; text mining.

Reference to this paper should be made as follows: Chen, X., Guo, C.H., Chau, M. and Zhou, W.H. (2013) 'Character usage in Chinese Short Message Service (SMS): a real-world study in Mainland China', *Int. J. Mobile Communications*, Vol. 11, No. 5, pp.429–445.

Biographical notes: Xi Chen is an Associate Professor of Information Systems at the School of Management, Zhejiang University, China. He obtained his BS (MIS) from Fudan University (China), MS (IS) from National University of Singapore and PhD (IS) from the University of Hong Kong. His research interests are in the areas of data mining and mobile services. His research has appeared or is forthcoming in *Decision Support Systems*, *European Journal of Operational Research*, *Journal of Organisational Computing and Electronic Commerce*, *Journal of the American Society of Information Science and Technology* and *Electronic Commerce Research and Applications*, and in the proceedings of several international conferences.

Chenhui Guo is currently a PhD Student in the Department of MIS, Eller College of Management, The University of Arizona. He obtained his Bachelor of Business Administration from Zhejiang University, Hangzhou, China. His current research interests include electronic markets, crowdsourcing and social media. He is also interested in research topics such as business intelligence and data mining.

Michael Chau is an Associate Professor in the School of Business at the University of Hong Kong. He received a PhD in Management Information Systems from the University of Arizona and a BSc in Computer Science and Information Systems from the University of Hong Kong. His research has appeared in top journals such as *ACM TMIS*, *ACM TOIS*, *CACM*, *DSS*, *IEEE Computer*, *IEEE TKDE*, *JAIS*, *JASIST*, *JMIS* and *MISQ*. He has been ranked as the 14th most productive researcher in the field of information science in 1998–2007.

Weihua Zhou is a Professor at the School of Management, Zhejiang University, China. He received his MS in Applied Mathematics from Zhejiang University, China, and PhD from the Department of Industrial Engineering and Logistic Management, Hong Kong University of Science and Technology. His research interests include supply chain management. His research has appeared in *Operations Research*, *Naval Research Logistics* and *Transportation Research Part B*: *Methodological*.

1 Introduction

Short Message Service (SMS) is one of the most commonly used electronically mediated communication service components of the mobile communication systems and now

available on most popular mobile networks (Constantinos-Vasilios and Ifigeneia, 2008). This popular service allows customer to send and receive short textual message, which is limited to 160 characters (or 70 characters in Chinese) each, instantly to and from other mobile phones in the digital cellular network.

Along with the fast adoption of mobile service throughout the world, SMS becomes a critical part of mobile service together with voice call. Because of keen competition upon voice service, telecommunication industry suffered from a steady decline in Average Revenue Per User (ARPU) in past years. To get revenue back, many telecommunication companies paid attention to the non-voice Value Added Services (VASs), such as SMS and instant message (Turel et al., 2007). It was estimated that 8 trillion text messages will be sent in 2011 (Mobithinking.com, 2011). By 2012, this number will reach 10 trillion (Xceedagents.com, 2011). Although consumers also use other types of messaging services such as e-mail and IM, usage of SMS still occupies the majority portion of consumers' messaging behaviour (Mobithinking.com, 2011).

Generally speaking, the fast growth in text messaging all around the world partially has resulted from relatively high value of Voice-to-SMS price ratio (Poulbere, 2005). SMS in countries with high Voice-to-SMS ratio (around 7 : 1), such as China and India, grow faster than countries with low Voice-to-SMS ratio (around 2 : 1). In China, the typical price for a short message is 0.1 RMB, which is relatively cheap compared with voice call (Hui, 2008). Besides, mobile commerce as a novel business model has drawn the attention of the IT world since the early 21st century. Previous study recognised SMS as a critical medium for mobile marketing, among which many marketing activities are based on SMS (Dickinger and Haghirian, 2004; Arpita and Sapna, 2012).

Previous studies on user behaviour in mobile service mostly depend on survey or experiment data. Differently, we choose actual short message data from real-world application, since we believe that the relevance of textual contents of SMS to the users' preferences is of great importance to the success of mobile advertising (Rau et al., 2011). By analysing textual contents of SMS messages, they can explore customer behaviour and preferences. SMS together with mobile instant message (IM) is a major data source of mobile customers, based on which analysts are enables to store, process and visualise customer behaviour in the mobile business environment.

We initiated the analytical work by using a computational approach that is commonly used in Chinese information retrieval. In this paper, we were engaged in extracting popular phrases used in Chinese short messages, based on Chien's PAT-tree approach (Chien, 1994, 1997). Through this work, we listed most popularly used bigrams, trigrams, 4-grams and 5-grams in Chinese SMS messages, thus analysed the phrase usage at a fundamental level. This forms the basis for future work on exploring popular topics in short messages by using more sophisticated text mining and information retrieval techniques. The short messages we used in the research were originally collected from two main state-owned telecommunication companies in Mainland China. The data source is quite unique and of great value for analysing the linguistic characteristics of Chinese short messages.

In Section 2, we present an in-depth review of current literature about phrase usage in SMS and PAT-tree approach. Section 3 concentrates on data collection and analytical methods, while Section 4 presents our findings. In Section 5, we deliver some discussion about the importance of our findings. Finally, the paper ends with conclusions and future directions in Section 6.

2 Literature review

2.1 Phrase usage in Short Message Service

Short messages are typically unstructured textual data, via which SMS users deliver meaningful information with each other. Automatic linguistic analysis of short messages can be achieved by using the state-of-the-art text-mining techniques. Text mining and information retrieval studies have been looking into a dozen of textural analytical tasks, including term extraction, text segmentation, text summarisation and so on. Also, they are now conducted on diverse data sources, such as business news (e.g., Ku et al., 2006), web pages (e.g., Qin et al., 2006; Zeng et al., 2011), technical patents (e.g., Tseng et al., 2007), financial statements (e.g., Ravisankar et al., 2011), e-mails (e.g., Zhang et al., 2004), blogs (e.g., Ku et al., 2006; Lu et al., 2010) and instant messages (e.g., Kucukyilmaz et al., 2008). However, little has been done about text mining on short messages. On the other hand, although keywords or feature extraction from text is always an important task in those text-mining studies, very few studies are devoted to analyse and describe the lexical patterns in text. A study on this topic is possible to explore interesting findings, such as users' preferences to products or services and their SMS usage behaviour. We believe that phrase-level–text-mining studies on short messages are of great potential for future research and business implementation.

Since mobile phones usually do not contain full-sized keyboard or powerful input instrument, short messages usually contain only a few words and brief meanings. Once invented, short message is regarded as a way to alert users or to deliver simple messages (Ling, 2005). Short messages are also commonly utilised as a method for greeting and reminding. Besides, commercial institutions may consider adopting SMS as a proper system for advertising and promotion (Rettie et al., 2005).

Studies concerning linguistic analysis of short messages seldom appeared in the previous literature. It might be due to the difficulty of collecting large-scale–real-world SMS message samples. To the best of our knowledge, we only found two studies about linguistic analyses on short messages. Both of them are mainly based on manual analyses, and the sample size of SMS messages they used is small. One of the works is Ling's study, which presented common message length, message complexity and frequently used words in SMS (Ling, 2005). This study summarised themes of short messages and listed several types of SMS contents. Those types listed are coordination, grooming, answers, questions, personal news and so on. Another related research showed that short messages are very different from textual contents of instant messages (IM) on PC platform (Ling and Baron, 2007). We believe with availability of large-scale data set of short messages and advanced computational linguistics approaches, more systematically and large-scale analyses of SMS messages should be conducted to reveal more findings.

2.2 Key-phrase extraction approach for free text in Chinese

Unlike English text, Chinese text cannot be segmented by spaces. Therefore, an approach to extract phrases from free text in Chinese is required to study usage of phrases in Chinese SMS.

There are three types of key phrase extraction approaches for free text in Chinese, namely dictionary approach, linguistic approach and statistical approach (Ong and Chen,

1999; Khoo et al., 2002). The dictionary approach runs fast. However, only phrases included in the dictionary can be extracted. In other words, great effort is needed to create a comprehensive dictionary and update it frequently to capture all possible usage of phrases. Usually, it is very difficult to achieve this goal. Linguistic approach extracts phrases based on syntactic and semantic rules. However, Chinese language system is rather complicated to code. Statistical approach has been shown to be efficient for Chinese text mining. It extracts phrases according to co-occurrence of characters or terms. Statistical approach is good at detecting new terms and names. Among a dozen of statistical approaches, Chien put forward a widely adopted phrase extraction method called PAT-tree Chinese phrase extraction approach (Chien, 1997, 1999). Once created, this approach has been adopted in several works in this area. Among those, Chau et al. (2007, 2009) utilised the PAT-tree approach in studying Chinese web-searching problems. They successfully extracted popular Chinese characters and terms in web searching. In their studies, major topics in web searching and the information-seeking behaviour of the users were analysed and presented.

Analysis on the text messages is different from web search log analysis, because SMS messages are usually in the format of free text. However, PAT-tree approach is still suitable for analysing Chinese SMS messages. We, therefore, adopted Chien's approach in SMS message analysis.

3 Data collection and analytical methods

The SMS data used in the current study were collected from three provinces (Hebei, Jiangxi and Zhejiang) in Mainland China. For Hebei Province and Jiangxi Province, the data were collected from China Mobile. For Zhejiang Province, the data were collected from China Unicom. We collected 13 samples from Hebei Province, six samples from Jiangxi Province and 12 samples from Zhejiang Province. Each sample was as large as 20 Megabytes and consisted of approximately 400,000 records.

The three provinces are in different areas of China and represent different levels of development. Zhejiang province is on the east coast and is one of the most developed provinces in China. Hebei province is in the north area of China and Jiangxi province is in the middle area of China. According to a recent study conducted by Baidu, the three provinces also represent different usage levels of mobile telecommunication. Zhejiang and Hebei are among the top 10 whereas Jiangxi's rank is 14 in terms of their citizen's usage of mobile telecommunication network (Baidu, 2011). The mobile users of Zhejiang, Hebei and Jiangxi account for around 5, 5 and 2.5%, respectively, of the total of China. So, the samples are believed to represent the general population of mobile phone users in China.

All the data was downloaded directly from the middleware of China Mobile and China Unicom. The commercial messages have already been filtered by the middleware. The phone numbers of the senders and receivers were also removed from the middleware to avoid the issue of privacy leak. We did not make any other changes to the original data to make sure that the data we analysed reflect the true situation.

The data was collected during the second week in August 2009. The time of data collection was evenly distributed across different time periods in a week as shown in Table 1. More samples were collected during peak times because there were more message activities during peak time according to our prior analysis.

Table 1 Time for data collection

	Weekdays	*Weekends*
Peak times (10.00 am–1.00 pm and 7.00 to 11.00 pm)	5 samples from Hebei and Zhejiang, 3 samples from Jiangxi	3 sample from Hebei and Zhejiang, 1 sample from Jiangxi
Non-Peak times	3 samples from Hebei and Zhejiang, 1 sample from Jiangxi	2 samples from Hebei, 1 samples from Zhejiang and Jiangxi

Most current study on mobile services usage or adoption is based on survey or experimental research. However, both of them may not be able to study the actual behaviours of users in real-world application. A recent trend is to analyse the actual behaviours of users through their usage data. However, a big challenge for such approach is that there are always too many data to analyse. Therefore, computational approach is used. We choose to use Chien's approach because it is one of the most popular approaches that have been used in different applications of Chinese free text analysis as we have discussed in Section 2.2.

The first step of Chien's approach is to build a PAT-tree, an indexing structure that has been used for free text processing both in Chinese and in English. The construction of PAT-tree relies on the concept of semi-infinite string. Basically, PAT-tree uses a binary tree to record the position of all possible semi-infinite strings in a text.

Not all the semi-infinite strings are phrases with complete meanings. The second step is to select Complete Lexical Patterns (CLPs) from all indexed semi-infinite strings. The selection of CLP relies on three measurements, namely association norm estimation (AE), Left Context Dependency (LCD) and Right Context Dependency (RCD) (Chien, 1997, 1999).

Suppose $x = x_1, x_2, x_3, \ldots, x_n$ is a lexical pattern under examination. Let y and z be the two longest substrings of x ($y = x_2, x_3, x_4, \ldots, x_n$). Association norm estimation is used to examine the association norm between y and z. The higher the value, the higher is the possibility that the y and z always appear together.

LCD measures how x is closely related to its adjacent left strings. Likewise, RCD measures how x is closely related to its adjacent right strings. Let $|L|$ be the number of unique left adjacent strings and $|R|$ be the number of unique right adjacent strings. Let α and β be the left adjacent strings and right adjacent strings of x, respectively. $|L|$ and $|R|$ should be bigger than a selected threshold t_1 and $\mathrm{Max}_{\beta} f(x\beta)/f(x)$ and $\mathrm{Max}_{\alpha} f(\alpha x)/f(x)$ should be smaller than another threshold t_2, otherwise x have left or RCD, where $f(x)$, $f(\alpha x)$ and $f(x\beta)$ are the occurrence frequency of x, x and its left adjacent string, x and its right adjacent string, respectively.

In summary, to identify a CLP, the following conditions (1), (2) and (3) must all be satisfied. The frequencies of phrases can be determined from the PAT-tree. The value of α and $|L|$ can also be extracted from an inverse PAT-tree, whereas the value of β and $|R|$ can be extracted from a normal PAT-tree. For details, please refer to Chien (1997, 1999).

$$T_1 = \mathrm{Min}\{|L|, |R|\} \geq t_1 \tag{1}$$

$$T_2 = \text{Max}\left\{\frac{\text{Max}_\alpha f(\alpha x)}{f(x)}, \frac{\text{Max}_\beta f(\beta x)}{f(x)}\right\} \le t_2 \tag{2}$$

$$T_3 = AE_x = \frac{f_x}{f_y + f_z - f_x} \ge t_3. \tag{3}$$

Clearly, t_1, t_2 and t_3 are three critical thresholds that help detect meaningful phrases extracted. However, little literature discussed the details on the determination of their values.

4 Results

After construction of the PAT-tree, we obtained all the possible semi-infinite strings in short message records. In Chinese language, each lexical pattern usually consists of 2–5 characters. Following this rule, the keywords we extracted from millions of short message records can be divided into four groups in terms of the length of the lexical patterns from 2 to 5 (i.e., bigrams, trigrams, 4-grams and 5-grams).

On the basis of our observations, we set value of t_3 at 0.02 for bigrams, 0.03 for trigrams, 0.3 for 4-grams and 0.4 for 5-grams. The threshold values and the number of phrases extracted are shown in Table 2.

Table 2 Threshold values for the three conditions

n-grams	*Bigrams*	*Trigrams*	*4-grams*	*5-grams*
t_1 condition	≥20	≥20	≥20	≥20
t_2 condition	≤0.3	≤0.3	≤0.3	≤0.3
t_3 condition	≥0.02	≥0.03	≥0.3	≥0.4
Number of phrases extracted	3861	4733	1932	601

4.1 Popular phrases in Chinese Short Message Services

To know the usage of Chinese phrases in SMSs, we look at the most popular bigrams, trigrams, 4-grams and 5-grams extracted from the records. Table 3 illustrates the top 50 popular bigrams and trigrams along with its English translation and frequencies in our data.

According to Table 3, we noticed that terms with two Chinese characters used in SMS are mostly about people's daily life. The term '今天' (today) is the most popular word in short messages, with a high average frequency of 19,069 among all samples. '明天' (tomorrow) is ranked second with an average frequency of 16,206. Including other top popular keywords such as '**晚上**' (night) and '下午' (afternoon), we found that a dozen of top popular bigrams in short messages are about time and schedule. People are more likely to use SMS to deliver information about their everyday life schedules. Besides, the frequent usage of terms such as '回来' (come back), '吃**饭**' (eat), '上班' (go to work) and '睡**觉**' (sleep) give us a brief description of common SMS users' daily activities. Living in a fast-developing country, Chinese SMS customers make good use of

this new technology to keep up with the fast pace of modern life. Another type of frequently used bigrams includes salutations between husbands (老公) and wives (老婆), indicating Chinese SMS customers most often send messages to their family members. Interestingly, the high occurrences of phrases '信息' (message) and '手机' (mobile phone) show that SMSs or other types of mobile services are hot topics in SMS. In the list of top bigrams used in Chinese SMS, most phrases are content words, but there are still several non-content-bearing phrases such as '一个' (one thing) and '也不' (neither). Different from character usage in Chinese web search engines revealed by Chau et al. (2009), phrase usage in SMS is more close to online free text.

The commonly used trigrams in Chinese SMSs are similar to bigrams. The most popular trigram is '**发信息**' (send message) with an average frequency of 2725. Information about people's daily activities takes up a large portion of the list, including '看**电视**' (watch TV) and '睡不着' (cannot fall asleep). We noticed that even the most popular trigram has a much lower frequency compared with those popular bigrams. In natural Chinese language, most meaningful terms are a single Chinese character or a combination of two Chinese characters. As a result, most terms with more than two characters in natural Chinese language are phrases built from two meaningful terms combined together. For instance, the popular trigram '看**电视**' (watch TV) is a combination of two terms, namely '看' (watch) and '**电视**' (TV). In addition, the term '石家庄' (Shijiazhuang, name of the capital city of Hebei Province) ranked 4th in the list of most popular trigrams, because we collected part of the research data from Hebei Province. Another interesting finding is that '人民**币**' (RMB), a business-related content term, is a heavily used phrase. As mentioned before, the bigrams and trigrams, which are most commonly used in Chinese SMSs, are mostly about people's daily life and are more close to free text in natural Chinese language.

However, unlike usage of bigrams and trigrams, usage of 4-grams and 5-grams, which are terms with more than three Chinese characters, has distinctive characteristics. In Table 4, we illustrated the top 50 popular 4-grams and 5-grams used in Chinese SMS (see Table 4).

An observation from the 4-grams and 5-grams in the phrases extracted from Chinese short message records reveals that terms with more than three characters are more related to commercial information, such as bank account and payment. The names of Construction Bank of China (CBC) and Industrial and Commercial Bank of China (ICBC) appear in the list as one of the most popular 4-grams. These two banks have huge service networks in the Mainland China and are involved in Chinese people's economic lives. The list of the most popular 4-grams and 5-grams also contains many bank-related terms such as '活期余**额**' (savings deposit balance) and '**储蓄**卡**账户**' (saving card account). There are other types of commercial services information as well, for example, '**详情请咨询**' (for further details please contact), '店**庆周年答**' (store's anniversary celebration), etc.

There are many conversational expressions in 4-grams and 5-grams. For example, there are also some *n*-grams expressing the meaning of grooming and greetings that appeared in the top 50 4-grams, such as '注意身体' (take care of your health), '好好休息' (take a good rest), '生日快**乐**' (happy birthday), '注意安全' (be careful), '做个好梦' (have a nice dream), '不要担心' (do not worry), '天天快**乐**' (happy everyday), '休息一下' (take a rest) and '好好学**习**' (study hard). In the top 50 5-grams,

phrases like '多穿点衣服' (wear warmly), '天天好心情' (in a good mood everyday), '好好睡一**觉**' (have a good sleep) and '**开心每一天**' (happy everyday) are very common greetings in Chinese, '到**时候再说**' (wait until then), '有**时间过来**' (come when available), '收到**请回复**' (please reply if receiving this message) and '以后有机会' (we will have a chance to do something in the future) are very common conversational phrases.

In summary, daily life is still a hot topic in 4-grams and 5-grams. Compared with bigrams and trigrams, 4-grams and 5-grams in Chinese SMS can express more complete meaning. Senders usually input these phrases to tell the recipient what to do, what the senders do and even describe the status of something briefly.

Table 3 Top 50 popular bigrams and trigrams used in Chinese SMS

Rank	*Bigram*	*Translation*	*Freq.*	*Trigram*	*Translation*	*Freq.*
1	今天	Today	19,069	**发信息**	Send message	2725
2	明天	Tomorrow	16,206	有**时间**	Have time	2691
3	回来	Come back	12,838	看**电视**	Watch TV	1809
4	老婆	Wife	11,983	**发短信**	Send short message	1748
5	**时候**	Time	11,335	早点睡	Sleep early	1506
6	吃**饭**	Eat	10,297	睡不着	Cannot fall asleep	1506
7	老公	Husband	10,138	一定要	Must (incomplete)	1504
8	**晚上**	Night	9799	好不好	is it ok?	1220
9	不好	Not good	9482	没**办法**	No way to do	1164
10	信息	Message	9334	人民**币**	RMB (Chinese currency)	994
11	**时间**	Time	9256	**发过来**	Send message to (Sb.)	969
12	手机	Mobile phone	9071	到**时候**	At that time	939
13	上班	Go to work	8784	不方便	Not convenient	915
14	一个	One (thing)	8695	没**关系**	Never mind	902
15	睡**觉**	Sleep	8326	一**辈子**	Lifetime	825
16	回家	Go home	8278	没**时间**	Have no time	810
17	**过**来	Come	8163	**过几天**	In a few days	793
18	一下	One time, a while	7952	明天去	Will do tomorrow	787
19	朋友	Friend	7704	不舒服	Be sick	779
20	也不	Neither	7175	个小**时**	(a few) Hours	774
21	下午	Afternoon	7039	明天早	Tomorrow morning (incomplete)	770

Table 3 Top 50 popular bigrams and trigrams used in Chinese SMS (continued)

Rank	*Bigram*	*Translation*	*Freq.*	*Trigram*	*Translation*	*Freq.*
22	一起	Together	6770	石家庄	(Capital of Hebei)	770
23	不想	Don't want to	6750	段时间	(a few) periods of time	750
24	回去	Come	6246	不认识	Don't know (somebody)	747
25	没事	Nothing	6071	女朋友	Girlfriend	726
26	告诉	Tell	5959	有一个	Have a..	724
27	短信	Short message	5948	今天不	Can't do today	716
28	还有	Still have (something)	5751	问一下	Ask about…	713
29	好好	Well	5646	办公室	Office	696
30	还没	Not yet	5620	一起去	Go together	693
31	休息	Take a rest	5609	开玩笑	Kidding	687
32	以后	Later	5468	一会儿	A while	680
33	宝贝	Honey	5240	明天上	Tomorrow morning (Incomplete)	668
34	昨天	Yesterday	5126	男朋友	Boyfriend	662
35	下班	After work	4849	一个月	One month	659
36	几天	Several Days	4713	身份证	ID card	641
37	因为	Because	4438	还以为	(Somebody) believe	637
38	有事	Busy	4434	回信息	Reply message	635
39	都不	Not	4354	长时间	Long time	619
40	亲爱	Darling	4231	等一下	Wait a minute	610
41	不用	Needn't	4204	也不想	Don't want too	607
42	一定	Must	4165	每天都	Every day (do something)	603
43	一点	A little	4154	打过来	Call me	589
44	家里	At Home	4093	今天下	This afternoon (Incomplete)	589
45	不回	Not go back	4022	不说话	Keep silence	587
46	工作	Work	4021	不相信	Not believe	579
47	已经	Already	3993	手机号	Cell number	574
48	东西	Something	3962	信息给	Message to (sb.)	573
49	开心	Happy	3896	不会来	Won't come	571
50	公司	Company	3771	呜呜呜	(sound of weeping)	566

Table 4 Top 50 popular 4-grams and 5-grams used in Chinese SMS

Rank	*4-gram*	*Translation*	*Freq.*	*5-gram*	*Translation*	*Freq.*
1	不好意思	I'm sorry	2822	账户余额为	The balance of account is…	608
2	早点休息	Rest early	1059	储蓄卡账户	Saving card account	490
3	建设银行	CBC (name of a bank)	886	阅读本邮件	Read this mail	370
4	注意身体	Take care of your health	815	分缴费成功	Successful payment instalment	342
5	时候回来	Come back at…	685	可直接回复	Reply message directly	303
6	好好休息	Take a good rest	668	缴费金额为	Payment amount is…	200
7	已成功为	Successfully	631	时候有时间	Have time at…	96
8	活期余额	Demand deposit balance	610	送多重好礼	Send gifts to	85
9	生日快乐	Happy Birthday	576	详情请咨询	For further details please contact	82
10	今天下午	Today afternoon	560	手机快没电	Cell phone out of battery soon	77
11	手机没电	Cell phone out of battery	543	手机歌曲库	Cell phone music storage	75
12	明天早上	Tomorrow morning	515	牡丹贷记卡	(name of a debit card)	69
13	温馨提示	Kindly remind	513	今天没上班	Didn't go to work today	66
14	心情不好	In a bad mood	501	牡丹信用卡	(name of a credit card)	64
15	注意安全	Be careful	497	今晚到明天	Overnight	64
16	有限公司	Limited corporation	494	信息也不回	Even not reply message	63
17	工商银行	ICBC (name of a bank)	477	多穿点衣服	Wear warmly	63
18	今天晚上	Tonight	444	店庆周年答	The store's anniversary celebration	62
19	明天上午	Tomorrow morning	438	到时候再说	Wait until then	61
20	呜呜呜呜	(sound of weeping)	426	酒店等服务	Hotel and other services	61
21	昨天晚上	Yesterday night	365	世界精彩尽	A wonderful world	61
22	发个信息	Send a message	349	月宽带时长	Monthly usage time for broadband	61
23	明天下午	Tomorrow afternoon	304	身体不舒服	Be sick	59
24	工行信使	Bank messager	284	天天好心情	In a good mood everyday	58
25	不回信息	Not reply message	273	龙卡信用卡	(name of a credit card)	57

Table 4 Top 50 popular 4-grams and 5-grams used in Chinese SMS (continued)

Rank	*4-gram*	*Translation*	*Freq.*	*5-gram*	*Translation*	*Freq.*
26	一点都不	Not at all	272	好好睡一觉	Have a good sleep	56
27	一起吃饭	Eat together	252	支出人民币	Pay RMB	55
28	好好照顾	Take good care of (somebody)	252	有点不舒服	Be a little sick	49
29	发个短信	Send a short message	242	中信信用卡	(name of a credit card)	48
30	不用担心	Don't worry	238	心里不舒服	Feel uneasy	48
31	做个好梦	Have a nice dream	234	一定要好好	Must do sth. well	47
32	回来吃饭	Come back for dinner	227	下午有时间	Have time in the afternoon	46
33	几点下班	When will be off work	216	有时间过来	Come when available	46
34	回家吃饭	Come home for dinner	215	长途话费多	Fee for long distance call	45
35	身体健康	Good health	206	网用户拨打	Users may call…	45
36	明天中午	Tomorrow noon	205	信息都不回	Even not reply message	44
37	明天还要	Will still do tomorrow	197	开心每一天	Happy everyday	43
38	一路顺风	Godspeed	193	账户卡取额	Account balance	43
39	今天中午	Today noon	190	收到请回复	Please reply if receiving (this message)	43
40	好好学习	Study hard	187	对身体不好	Not healthy	43
41	建行对账	CBC (a bank) balance	187	账户折取额	Account limit	41
42	保重身体	Take care	184	世界好友周	International Friendship Week	40
43	天天快乐	Happy everyday	183	款越野吉普	(a model of) Cross-county jeep	39
44	账户余额	Account balance	182	以后有机会	(we) will have a chance to do something in the future	39
45	休息一下	Take a rest	171	北风成焦点	North wind becomes the focus	39
46	友情提示	Friendly remind	170	降雨不明显	Raining slightly	39
47	即日起至	From now on	167	明天要上班	Have to work tomorrow	38
48	工作顺利	Hope your work is going well	166	说话不算数	break one's promise	36
49	好想好想	Missing (somebody) very much	162	体验海滨酒	Experiencing a brand of wine	33
50	时候过来	Come back at…	160	发信息过来	Send message to me	33

4.2 The Zipf distribution of the extracted phrases in SMS

The Zipf law states the relationship between frequencies and ranks of word tokens in a corpus. The widely used form of Zipf distribution is as follows (Mandelbrot, 1953)

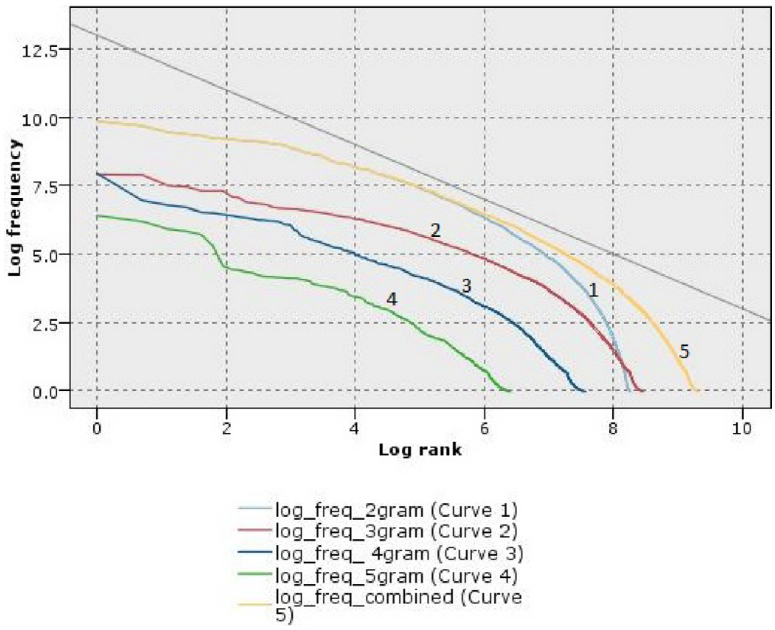
$$f = k / (\gamma + \alpha)^{\beta} \tag{4}$$

where α and β are constants for a text being analysed. The Zipf curve shows the linear relation between log-rank and log-frequency. According to the literatures, Zipf curve should have a slope of −1 if the data follow the Zipf distribution.

Many studies have shown evidence that distribution of word tokens in various corpuses follows Zipf distribution. Ha et al. (2003) examined distribution of characters in both English and Chinese. They found that combined Zipf curve of bigrams, trigrams and 4-grams complied to Zipf's law approximately. Spink et al. (2001) used a double-log rank-frequency plot to show that Excite search log data followed the Zipf distribution. Chau et al. (2009) also discovered that characters usage in Chinese web search log followed Zipf distribution.

In our research, we drew Zipf curves of all four categories of grams and all *n*-grams combined in Figure 1, and conducted regression analyses to calculate the value of slopes. As shown in Figure 1 (where the straight line is '$y = 13 - x$'), the curves fall gently at the beginning and went down horizontally with the straight line $y = 13 - x$. Zipf curve of the combined phrases is at the top of the figure, followed by the curves of bigrams, trigrams and 4-grams. The curve of 5-grams is at the bottom of the figure, reflecting the situation that more bigrams and trigrams are extracted from the SMS records than 4-grams and 5-grams. In contrast, fewer popular 4-grams and 5-grams can be found in short messages.

Figure 1 Zipf curve for extracted phrases in Chinese SMS (see online version for colours)



To know precisely about the relationship between natural logarithm of ranks and natural logarithm of frequencies, we presented the results of Ordinary Least Square (OLS) estimation of linear regression analyses with a slope and an intercept for the *n*-grams in Table 5. As shown in the table, the Zipf curves for the four categories of extracted phrases have negative slopes (coefficients) ranging from −1.251 to −1.901. The coefficient of Zipf curve of all *n*-grams combined is −1.847. The *R* squares of the regression models are very close to 1.0, showing that the frequencies of extracted terms follow Zipf distribution. In general, curve graph and result of regression analysis verify that phrases extracted from Chinese short message records fit Zipf distribution quite well.

Table 5 The results of regression analyses of double-log plots

n-grams	*Intercept* (*Std. Error*)	*Sloop* (*Std. Error*)	*R Square*
Bigrams	17.448***	–1.901***	0.840
	(0.098)	(0.013)	
Trigrams	11.155***	–1.586***	0.879
	(0.064)	(0.009)	
4-grams	11.324***	–1.446***	0.924
	(0.063)	(0.009)	
5-grams	8.254***	–1.251***	0.950
	(0.064)	(0.012)	
All n-grams	18.010***	–1.847***	0.904
	(0.048)	(0.006)	

Significance level: ****p*-value < 0.001.

5 Discussion

5.1 Academic implications

A study on language usage characteristics of SMSs is of great interests but relatively rare. Ling (2005) conducted a sociolinguistics of SMS analysis of Norwegians. A similar research was carried out in 2007, which compared the linguistic characteristics of SMS and online instant messaging services in English (Ling and Baron, 2007). However, very few studies have focused on SMS in Chinese. Our study revealed some interesting findings regarding characters usage patterns in Chinese SMS.

- Characters usage pattern in Chinese short message followed Zipf distribution. However, the coefficients of bigrams, trigrams, 4-grams and 5-grams showed a different pattern from those reported in web search log analysis (Chau et al., 2009). In our study, bigrams have the lowest coefficient and 5-grams have the highest coefficient, which indicates that the usage of bigrams was more diversified than trigrams, 4-grams and 5-grams in SMS. On the other hand, Ha et al. (2003) reported that the coefficients in Chinese free text were close to −1. In our study, the coefficients for bigrams, trigrams, 4-grams and 5-grams are very close to −1. In comparison, in the web search log analysis (Chau et al., 2009), the coefficients of

trigrams, 4-grams and 5-grams were different from −1. We may conclude that the language of SMS in Chinese is more similar to language in free text in Chinese.

- Compared with the studies on Chinese free text analysis (Ha et al., 2003) and web search log analysis where most frequently used bigrams are nouns about things, in our study most bigrams are about time, personal pronouns, or names. This indicates that there are more conversational elements in the language of SMS than topic elements.
- From Tables 2 and 3, it can be observed that frequency of bigrams is much higher than those of trigrams, 4-grams and 5-grams, which means that basic units of SMS language are bigrams. This phenomenon may be due to the fact that typing in hand phone is not an easy task and length of SMS is limited to 70 Chinese characters, so that people tend to send short and quick information to others.

5.2 Managerial implications

Our results are also meaningful to mobile service providers. The key words analysis reveals that issues about bank account and daily life are the two most popular topics in SMS communication. This fact reflects two challenges to current operators of mobile services. First, protection of privacy information is a crucial problem. Service providers should pay much more attention to protecting privacy information to avoid safety issues such as phishing attacks via SMS services. The second challenge is how to make use of SMS information legally. Detailed analysis of SMS information can reveal important preferences of mobile phone users, for example, what a people may need when and where. Making use of those context information can help provide more sophisticated and personalised services to users. However, it is always important to make users feel that their privacy is protected and has not been misused.

6 Conclusion and future work

In this paper, we reported an exploratory study of Chinese *n*-grams usage in SMS. We analysed SMS logs from three different provinces of China. To our knowledge, this paper provides detailed analyses and discussions on Chinese *n*-gram usage in SMS that were not reported in previous studies in this area. Our analysis consisted of three different aspects. First, we studied distribution of LCD, RCD and Association Norm (AN) in our sample. The findings in our study can provide some clues for the determination of the threshold values for the extraction. Second, we studied the most popular *n*-grams in Chinese SMS text. We discovered in this study that bigrams and trigrams in Chinese SMS are mostly related to daily life, such as time and activities. In contrast, 4- and 5-grams contain more complete semantic meanings and, therefore, are used to express greetings, grooming and commercial information to the receivers.

Note that the data collected in the current study was pre-processed by the middleware, which prevented privacy disclosure before we downloaded the logs to our computers. Because of the issue of protection, all the telephone numbers were removed and all the records were combined. As a result, the frequencies reported in this paper are averages over all the samples. But considering the size of the sample, it is still acceptable.

On the other hand, the current data came from three major provinces in China, owing to differences in culture and dialects, the findings may not be directly applicable to other Chinese users (e.g., Hong Kong and Taiwan). Future research will be needed to study the similarities and differences in the *n*-gram usage characteristics among these groups of users.

Acknowledgements

The authors thank the Editor-in-Chief and the reviewers for helpful comments on earlier drafts of this paper. The research related to this paper was directly supported by a research grant from the National Natural Science Foundation of China (Project No. 71002083 and 71072118), a research grant supported by the PhD Programme Foundation of the Ministry of Education of China (Project No. 20100101120096), and a key research project of National Social Science Foundation (Project No. 10&ZD047).

References

Arpita, K. and Sapna, R. (2012) 'Mobile marketing in Indian retail: a preliminary investigation of relationship and promotional endeavours through Short Message Service', *International. Journal of Business Competition and Growth*, Vol. 2, No. 2, pp.110–128.

Baidu (2011) *A Report on the Development of Chinese Mobile Internet*, Obtained from the Internet: http://developer.baidu.com/download?f=2011Q3 (Accessed July, 2012).

Chau, M., Fang, X. and Yang, C.C. (2007) 'Web searching in Chinese: a study of a search engine in Hong Kong', *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 7, pp.1044–1054.

Chau, M., Lu, Y., Fang, X. and Yang, C.C. (2009) 'Characteristics of character usage in Chinese web searching', *Information Processing and Management*, Vol. 45, pp.115–130.

Chien, L-F. (1994) 'PAT-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval', *Information Processing and Management*, Vol. 35, No. 4, pp.501–521.

Chien, L-F. (1997) 'PAT-tree-based keyword extraction for Chinese information retrieval', *Proceedings of the ACM SIGIR International Conference on Information Retrieval (SIGIR '97)*, Philadelphia, PA, USA, pp.50–59.

Chien, L-F. (1999) 'PAT-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval', *Information Processing and Management*, Vol. 35, No. 4, pp.501–521.

Constantinos-Vasilios, P. and Ifigeneia M. (2008) 'Mobile services: potentiality of Short Message Service as new business communication tool in attracting consumers', International Journal of Mobile Communications, Vol. 6, No. 4, pp.456–466.

Dickinger, A. and Haghirian, P. (2004) 'An investigation and conceptual model of SMS marketing', *Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS '2004)*, pp.1–10.

Ha, L.Q., Sicilia-Garcis, E.I., Ming, J. and Smith, F.J. (2003) 'Extension of Zipf's law to word and character N-grams for English and Chinese', *Computational Linguistics and Chinese Language Processing*, Vol. 8, No. 1, pp.77–102.

Hui, L. (2008) 'Boss boosts China mobile Shandong', *Huawei Technologies Magazine*, Vol. 40, pp.38–40.

Khoo, C.S.G., Dai, Y. and Loh, T.E. (2002) 'Using statistical and contextual information to identify two- and three- character words in Chinese text', *Journal of the American Society for Information Science and Technology*, Vol. 53, No. 5, pp.365–377.

Ku, L-W., Liang, Y-T. and Chen, H-H. (2006) 'Opinion extraction, summarization and tracking in news and blog Corpra', *Proceedings of AAAI Symposium on Computational Approaches to Analysing Webglogs 2006*, pp.100–107.

Kucukyilmaz, T., Cambazoglu, B.B., Aykanat, C. and Can, F. (2008) 'Chat mining: predicting user and message attributes in computer-mediated communication', *Information Processing and Management*, Vol. 44, No. 4, pp.1448–1466.

Ling, R. and Baron, N.S. (2007) 'Text messaging and IM: linguistic comparison of American college data', *Journal of Language and Social Psychology*, Vol. 26, No. 3, pp.291–298.

Ling, R. (2005) 'The sociolinguistics of SMS: an analysis of SMS use by a random sample of Norwegians', in Ling, R. and Pederson, P.E. (Eds.) *Mobile Communications: Re-negotiation of the Social Sphere*, Springer, London, UK, pp.1431–1496.

Lu, L., Zhu, F., and Hu, B. (2010) 'A novel method to detect latent community in Blogspace', *Journal of Computational Information Systems*, Vol. 6, No. 7, pp.2151–2157.

Mandelbrot, B.B. (1953) 'An information theory of the statistical structure of language', in Jackson, W. (Ed.): *Communication Theory*, Academic Press, New York, pp.503–512.

Mobithinking.com (2011) *Global Mobile Statistics 2011: All Quality Mobile Marketing Research, Mobile Web Stats, Subscribers, Ad, Revenue, Usage, Trends…*, Obtained from the Internet: http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats (Accessed August, 2011).

Ong, T-H. and Chen, H. (1999) 'Updateable PAT-tree approach to Chinese key phrase extraction using mutual information: a linguistic foundation for knowledge management', *Proceedings of the Second Asian Digital Library Conference*, pp.63–84.

Poulbere, V. (2005) 'The story of SMS global market development', *Huawei Technologies Magazine*, Vol. 17, pp.39–43.

Qin, J., Zhou, Y., Chau, M. and Chen, H. (2006) 'Multilingual web retrieval: an experiment in English-Chinese business intelligence', *Journal of the American Society for Information Science and Technology*, Vol. 57, No. 5, pp.671–683.

Rau, P.P., Zhang, T., Shang, X. and Zhou, J. (2011) 'Content relevance and delivery time of SMS advertising', *International Journal of Mobile Communications*, Vol. 9, No. 1, pp.19–38.

Ravisankar, P., Ravi, V., Rao, G.R. and Bose, I. (2011) 'Detection of financial statement fraud and feature selection using data mining techniques', *Decision Support Systems*, Vol. 50, No. 2, pp.491–500.

Rettie, R., Grandcolas, U. and Deakins, B. (2005) 'Text message advertising: response rates and branding effects', *Journal of Targeting, Measurement and Analysis for Marketing*, Vol. 13, No. 4, pp.304–312.

Spink, A., Wolfram, D., Jansen, B.J. and Saracevic, T. (2001) 'Searching the web: the public and their queries', *Journal of the American Society for Information Science and Technology*, Vol. 52, No. 3, pp.226–234.

Tseng, Y-H., Lin, C-J. and Lin, Y-L. (2007) 'Text mining techniques for patent analysis', *Information Processing and Management*, Vol. 43, No. 5, pp.1216–1247.

Turel, O., Serenko, A. and Bontis, N. (2007) 'User acceptance of wireless short messaging services: deconstructing perceived value', *Information and Management*, Vol. 44, pp.63–73.

Xceedagents.com (2011) *SMS Market – Global Perspective*, Obtained from the Internet: http://www.xceedagents.com/images/research-assistant/web-research/SMS-Market-Global-Overview-Task-Update.doc (Accessed August, 2011).

Zeng, D., Wei, D., Chau, M. and Wang, F. (2011) 'Domain-specific Chinese word segmentation using suffix tree and mutual information', *Information Systems Frontiers*, Vol. 13, pp.115–125.

Zhang, L., Zhu, J. and Yao, T. (2004) 'An evaluation of statistical spam filtering techniques', *ACM Transactions on Asian Language Information Processing*, Vol. 3, No. 4, pp.243–269.