Statistics and Analysis of Mongolian Syllables Based on Network Corpus

Zhuyuan Cai[1] and Monghjaya[2,*]
College of Computer Science, University of Inner Mongolia, Hohhot, Inner 010010, China.
*Corresponding author

***Abstract*—This article achieved the large-scale Mongolian text corpus from CCTV and some other news websites, and conducted statistics and analysis on the Mongolian syllables in this text. From the statistics and analysis, we can see that the possibility of the co-occurrence of the different Mongolian syllable by the n-gram model. At the same time, these data also show that the main reasons leading to the misspelling of Mongolian include the following aspects: one is the monosyllabic error, the second is the misuse of the space, the third is the improper use of the control character, and the fourth is the polyphonic word of the same shape.**

Keywords—mongolian syllable; n-gram model; spell check; statistics and analysis; network corpus

I. Introduction

Text proofreading is one of the main application fields of natural language processing. In recent years, many scholars have tried and studied the project from different aspects and angles. These studies apply different methods to text proofreading and also build a knowledge base about the rules of Mongolian. However, there are no published article on the misspelling of the Mongolian webpage texts. In this paper, we will analyze the misspellings of large-scale network Mongolian texts, and then to achieve the main reasons for the misspellings in Mongolian texts in real texts, so as to provide a basis for Mongolian text proofreading work, and at the same time, it would investigate Whether network Mongolian text can be used to build a high-quality Mongolian corpus.

Firstly, this article will introduce the status on the research of the Mongolian text proofreading at the present stage; Second, it would explain the methods and routines of obtaining the corpus in this paper and the specific conditions about the obtained corpus; then you would be told about the specific analysis result of the Mongolian syllables on this corpus, and finally we would summarize the full text.

II. Research Status

For the text proofreading of Mongolian, in 1997, Mr. Hua Shabao of Inner Mongolia University, a Mongolian automatic proofreading system MHAHP based on the additional components of Mongolian word formation, was mainly based on the Mongolian orthographic rules for text proofreading[1]. In 2007, Zhao Jun et al. proposed a 2-gram Mongolian proofreading model based on syllables, which is based on the influence of the combination of Mongolian vowels and consonants on the composition of syllables, and then the construction of a syllable library is used to judge the correctness of Mongolian words[2]. In 2009, S·Raglau proposed a Mongolian proofreading algorithm based on Non-Determinate Finite Automata[3]. This method greatly improved the speed of the algorithm and improved the automatic checking of non-word errors. Hao Li et al. used Bayesian algorithm to implement text correction for news corpus, and the experimental error correction rate was about 89%[4]. In 2013, Su Chuanjie et al. proposed a Mongolian proofreading model based on the statistical translation framework[5]. The error-correction rate of non-word errors reached 97.55% on its test set.

Most of the above researches are aimed at the construction of Mongolian text proofreading tools and the construction of related knowledge bases such as syllable rules. Few people have discussed and analyzed the situation of Mongolian syllables for large-scale real Mongolian corpora. This paper will conduct a statistical analysis of large-scale Mongolian corpus to examine the real situation of Mongolian syllables in real texts and to find out how to provide reliable support for the construction of Mongolian text corpus.

III. Data Acquisition and Processing

This section will discuss about the acquisition of large-scale Mongolian network texts, syllable segmentation and the basis of the syllable spelling.

A. Source of Corpus

Through careful screening, we obtained five news-based Mongolian websites from five Mongolian websites as the source of text corpus. The basic information of these five websites is shown in Table I. China Mongolian Broadcasting Network and Inner Mongolia Xinhua Net are based on international Standard coding, the rest are based on the Menksoft code, which need to be transcoded into an international standard code. In the process of transcoding, we used the Mongolian code conversion system developed by the Mongolian Information Processing Technology Key Laboratory of Inner Mongolia University. The system has been modified and used over and over, so the conversion rate is trustworthy. The information extracted from these pages. It only contains the Mongolian part of the body of the article, and the rest has been removed, thus providing a guarantee for the correct use of the corpus.

TABLE I. BASIC INFORMATION ABOUT THE WEBSITES

S/N	Domain Name	Website Name	Coding Scheme	Coder
1	cctv.com	mongol.cctv.com	UTF-8	Menksoft
2	mongolcnr.cn	www.mongolcnr.cn	UTF-8	Unicode
3	people.com.cn	mongol.people.com.cn	UTF-8	Menksoft
4	news.cn	mongolian.news.cn	UTF-8	Unicode
5	mglbbs.com	www.mglbbs.com	UTF-8	Menksoft

B. Corpus Acquisition Method

In this article, we mainly use the Beautiful Soup, regular expressions and the urllib packages to get the relevant information we need from the webpage. We analyzed the Mongolian webpage and found that the title, release time, and body of the article belong to different modules, and there are clear segmentation marks to distinguish them at the same time. When we need to obtain content, we only need to find the corresponding sections. Then use python's Beautiful Soup package to extract the corresponding content.

C. Syllable Segmentation Method

1) Syllabic segmentation rules and methods for Mongolian

The vowel is the primary element of a syllable, and there is usually only one vowel in a syllable. Consonant letters play an auxiliary role, generally do not constitute a syllable alone, and multiple consonants can appear at any position. When a combination vowel occurs, the former is a positive vowel and the latter is a negative vowel.

2) Syllable decomposition method [6]

a) Traverses the string, disconnecting from the consonant when there is a consonant between the two vowels

b) When there are two consonants in the middle of two vowels, disconnect from the middle of the two consonants

c) More broadly, when there are N consonants in the middle of the two vowels, they are disconnected from the last consonant.

According to these rules, the Mongolian text is easily cut into different syllables, and the frequency of occurrence of each Mongolian syllable can be counted. When we split these data, we find that there are lots of kinds of the Mongolian syllables in the data set. Some of the Mongolian syllables with higher frequency appearing in the above corpus are shown in Table II (In this table, ‘V’ represents vowel, ‘C’ represents consonant.). According to these data, on the one hand, we can see that a syllable can be formed alone with one Mongolian letter. at the same time, the frequency of occurrence is still relatively high. On the other hand, a syllable with a higher frequency is mainly composed of two or three phonemes. The specific frequency and rule can be achieved from the following table.

TABLE II. THE FREQUENCY OF DIFFERENT KINDS OF MONGOLIAN SYLLABLES IN NETWORK CORPUS

S/N	The Class of Syllables	Frequency(Times)	Rate (%)
1	CV	2507808	54.69
2	CVC	1382264	30.15
3	CVV	194384	4.24
4	V	172840	3.77
5	CC	170249	3.71
6	VC	99565	2.17
7	CVVC	28673	0.63
8	VV	11876	0.26
9	CCV	6451	0.14
10	CVCC	3269	0.07
11	VCV	2643	0.06
12	VVC	2434	0.05
13	CCVCC	869	0.02
15	CCVC	435	0.01
16	VCC	383	0.01
17	CCCC	261	0.01

D. Corpus Size

Using the above method, we obtained 23,840 Mongolian pages from the Internet, including 240,000 Mongolian, 2.33 million words, and 4.58 million syllables. Detailed statistics are shown in Table III .

TABLE III. THE SIZE OF THE ACQUIRED NETWORK CORPUS

S/N	Website	URLs	Words	Sentences	Syllables
1	mongol.cctv.com	3358	122074	13874	276832
2	www.mongolcnr.cn	9744	386054	34675	690648
3	mongol.people.com.cn	357	64670	7999	147574
4	mongolian.news.cn	7261	1474084	154054	3180989
5	www.mglbbs.com	3120	284660	29644	566132
Sum		23840	2331542	240246	4585343

IV. Statistics and Analysis of Mongolian Syllables

A. The Main Classifications and Reasons for the Misspelling of Mongolian Syllables

1) Extra letters

In the process of inputting Mongolian words, one or more letters or characters are inserted, resulting in syllabic segmentation errors, which lead to the error for the entire word.

2) Missing letters

In the process of inputting Mongolian words, missing letters will also lead to syllable segmentation errors, which will make the words impossible.

3) Wrong letters

Wrong letters are generally divided into two categories. One is misspelling which leads to that syllables can’t be formed.

The other error is caused by the Mongolian orthographic rules. The kind of error needs to be corrected and identified by the Mongolian orthographic rules. The specific rules are as follows:

a) If the beginning vowel of the word is a positive vowel, then the vowel in the middle and the ending of the word should be positive vowels; meanwhile, if the beginning vowel of the word is a negative vowel, t then the vowel in the middle and the ending of the word should be a negative vowel; in addition, if the first vowel is a neutral vowel, the first negative or positive vowel is used to determine the positive or negative. If it is totally neutral, the word would be considered a negative word[7].

b) The lip vowels 0 (ᠥ) and o (ᠣ) can only be used as the first vowel of the word, and the v(ᠦ) and u(ᠤ) should be in the middle of the word.

c) Consonant harmony rules

4) The vowel separator is replaced by a narrow width without spaces.

5) A control char*acter is entered multiple times in succession.*

6) One word has a variety of pronunciations [8].

B. Statistics and Analysis of Syllables

According to the previous analysis of Mongolian syllable data, we used the 2-gram model and the 3-gram model to obtain the co-occurrence frequency of Mongolian syllables based on the above syllables of Mongolian news network. This paper selects the syllables with higher co-occurrence frequency to research on it. From the statistical data, we can confirm that some syllables can co-occur together. For example, it is very likely that the Mongolian syllable '[illegible]' will appear after the Mongolian syllable '[illegible]'. This can be summarized from these data, and we can conclude that some syllables would not co-exist. The statistics of these rules will provide some help for the improvement of the accuracy of our future Mongolian text proofreading.

V. CONCLUSION

In this article, we obtained a total of 23,840 Mongolian pages from the web. According to these pages, we obtained about 240,000 Mongolian words, 2.33 million Mongolian words, and nearly 4.6 million syllables of Mongolian text corpus. Statistics and analysis were carried out. From the statistics and analysis data, we can see that the possibility of the co-occurrence of the different Mongolian syllable by the n-gram model**.** At the same time, we conclude that the main reasons leading to the misspelling of Mongolian include the following aspects: one is the monosyllabic error, the second is the misuse of the space, the third is the improper use of the control character, and the fourth is the word having a variety of pronunciations.

REFERENCES

[1] Hua Shabao. Modern Mongolian Automatic Proofreading System——MHAHP[J]. Journal of Inner Mongolia University: Philosophy and Social Sciences, 1997 (4): 49-53.

[2] Zhao Jun. Design and implementation of Mongolian lexical analysis corrector based on syllable statistical language model [D]. Inner Mongolia University, 2007.

[3] S·Raglau. Mongolian proofreading algorithm based on uncertain finite automata [J]. Journal of Chinese Information, 2009, 23(6): 110-115.

[4] Hao Li, Yan Dengbala, Gong Zheng, et al. Automatic proofreading of Mongolian text based on Bayesian algorithm [J]. Journal of Inner Mongolia University: Natural Science Edition, 2010, 41(4): 440-442.

[5] Su Chuanjie, Hou Hongxu, Yang Ping, et al. Mongolian automatic spelling proofing method based on statistical translation framework [J]. Journal of Chinese Information, 2013, 27(6): 175-179.

[6] Bai Shuangcheng, Hu Qitu, Mu Ren. Implementation and Application of Mongolian Syllable Segmentation Algorithm[J]. Research on National Language and Information Technology——Proceedings of the 11th National Symposium on National Language and Literature Information, 2007.

[7] Fulin Ga. Mongolian Orthography Research [M]. Hohhot. Inner Mongolia People's Publishing House. 2001.

[8] Monghjaya, Shan Dan. Implementation of Mongolian Transcoding to Latin Transfer and Split-Symbol Algorithm[J]. Chinese Journal of Information Science, 2011, 25(4): 101-104.