Summary of the Uyghur language information processing

Guixian Xu, Nuerguli, Qi Liu, Mengjuan Yang, Guanhao Feng, Sudian Wu
MinZu University of China, Beijing, China
xuguixian2000@sohu. com

Keywords: Uyghur information processing; Search engine; Public opinion supervision

Abstract. As the mother tongue of Uyghur，Uyghur language plays an important role in the expression and heritage of Uyghur The development of Uyghur information processing technology directly affects the degree of informationization and intellectualization of minority areas. This paper will summarize the current status of Uyghur information processing technology development, and present some deficiencies and suggestions for improvement.

Introduction

Modern Uyghur has 32 letters, including 8 vowels and 24 consonants. In xinjiang Uyghur Autonomous Region，Uyghur language(also known as the ancient Uyghur) that is similar with Arabic is regarded as the official language. It is widely used in network applications, communication. With the popularity of the windows system and the spreading of object-oriented programming ideas, Uyghur information processing has got an unprecedented development opportunities. Many colleges , universities and companies have developed their own Uyghur information processing systems.

The Meaning and the Value of Research on Uyghur Information Processing

Uighurs are the main ethnic minorities living in the XinJiang Uygur Autonomous Region, and they have their own language and scirpt. In the 21st century, if Uyghur language have not entered the network information age, carrying the Uighur culture and spirit, it could face the danger of being eliminated. In this case, the development of Uyghur information processing technology is an urgent need. There is a great vallue that the development and research on Uyghur information processing technology will greatly promote the development of science and technology in Uygur Autonomous Region.

The Current State of Uyghur Information Processing

Unicode is a character encoding used on the computer, which is also known as sigle code, IWC yars. We know that computers only deal with numbers, using numbers to store letters or other characters. Before the appearance of Unicode, there is no unified coding standard, which means that it will make the representation of information confusing. However, Unicode provides for any character in different countries a unique binary digital code, no matter what platform, what procedures, what language.

Unicode began to develop in 1990 and in 1994 it was officially announeced. In 1991, the International Organization for Standardization and the Unicode Consortium decided to work together to shape the Arabic text (Uyghur included) to develop a suitable right-to-left writing generic coding standard. Document No. 641 and No. 434 of ISO10636 international coding standard let the Uyghur language into the international character encoding standard. In 1994, Uyghur international standard was announced to the world on the Unicode meminar held in Turkey. Therefore, Uyghur alphabet successfully entered the international Unicode encoding standart, which laid the foundation for its further development. [1]

Microsoft office is a windows-based software developed by Microsoft. The software have been widely used in the whole world because of its easy-to-use and convenience. However, office does

not have the ability to completely deal with Uyghur information at the time. After office 2000 version, this software provides a kind of right-to-left writing direction to the public, and supports Unicode. Based on these new properties of office 2000, some scholars proposed a method for helping people to realize Uyghur information processing on the office. To know more details please refer to the original[2]. Thus Uyghur information processing on office 2000 had been realized and applied.

Not long after, applied Uyghur information processing on office, it was also found its shortcomings. Athough capable of handling Uyghur information, Microsoft office can not conduct Uyghur syllable line breaks and automatic strtch. By 2004, Uyghur language had been widely used in government offices and people's daily life. But there is no such a perfect software that can handle Uyghur information. So it had an vital effect on Uyghur Autonomous Region's technological development. For these cases, some scholars like LuYouFei , from Chinese Academy of Science , designed and developed Uyghur version of office which having the ability to deal with more alphabet problems. So the problem had been completely solved. The key technologies includes automatic selection shape, according to syllable line breaks and automatic stretch. [3]

In 1998, the first Uyghur website was established. Up to now, Uyghur language has become so common that the number of Uyghur netizens is contiunally increasing as well. In accordance with The Electronic Commerce Development Report of Xinjiang, 2014, released by the commerce department of Xinjiang Uyghur Autonomous Region, till 2013, Xinjiang had seen a breakthrough of its netizens over 10 million scale, reaching 10,940,000, while the penetration rate at 49%[4]. The Uyghur imformation and coments that Uyghur netizens can get, conveys the local culture and it also reflects network public sentiments. In this case, the Uyghur information processing based on web becoms neccessary.

When handling Uyghur Internet text information, we must first crawl the page, remove HTML notes . Chinese word segmentation has several basic method, such as Forward Maximum Matching Algorithm, Reverse Maximum Matching Algorithm and the Minimum Segmentation Algorithm[5]. Nevertheless, Uyghur segmentation processing technology is different from Chinese and English. Uyghur belongs to the altaic turkic and agglutinative language on the grammatical structure. Generally speaking, the word is the smallest unit in language processing. Uyghur words are separated by space, so, in some degree, there doesn't exist segmentation problems in Chinese processing[6]. But Uyghur words have abundant changes in morghology, new words can be structured by adding prefixes and postfixes, and meanwhile, various grammatical meaning are expressed by various prefixes and postfixes. Usually, the structure of Uyghur words can be expressed as, "prefix + word + postfix"[7]. Abundant changes in word morphology increses the difficulty of Uyghur segmentation.

In 2004, two professors, from Xinjiang University , proposed a method of Uyghur segmentation. The basic idea was: the establishment of a basic lexical rules and thesaurus, cut its substring for a given pending Participle the Uyghur Paragraph W in accordance with the provisions of the lexical rules , if the substring match a term of lexicon , remove the word from the lexicon lexical analyzed by lexical rules. Otherwise, continue to the next discourse segmentation. [8]

After the Uyghur word segmentation method became feasible, Uighur text classification and information retrieval could be used in the subsequent process of text classification and information retrieval. The vector space model, weight calculation formula, the similarity calculation formula, can be used to realize the clustering of Webpage text topic identification. Combined with the characteristics of Uighur language, Xinjiang University[9], Chinese Academy of Sciences, physical and Chemical Research Institute and other universities proposed their Uyghur information processing method based on Web information retrieval. According to the characteristics of Uighur itself, master of the Chinese Academy of Sciences, Li Yanjiao and Jiang Tonghai proposed the Bias classifier improved weight, constructed a Uighur text classification model. They put forward a combination of chi square test and document frequency weighted Bias classifier, and gived a kind of Uighur text classification model. Using the micro average accuracy and macro average F1 value as the evaluation standard of text classification system, the experimental verification were carried

out on the collected Uyghur corpus. Experiments showed that algorithm they proposed could have better classification results and better performance in Uighur text classification[10].

Before 2004, there were already a lot of numbers of Uyghur websites, but we didn't have any kind of Uyghur information search engine. As a result, Uighur users had no way to search Uyghur information on the Internet. Luckly in 2004, graduate from XinJiang university AiSaTiJang published his thesis presenting a design and implementation method of weaving of search engine. In his design method, by means of traversing, data acquisition, filtering process, it can eventually achieve the retrieval and query information. Due to this, the author had initially realized the Uyghur search engine, which is a great begining[11].

Some Deficiencies and Several Suggestions for Improvement

Uyghur encoding scheme has entered into the Unicode standard. But due to the lack of propaganda and communication, the international standard is ignored. It brings great inconvenience to the exchange of information that the Uyghur character encoding scheme hava hundreds of variety, and they are not compatible with each other. Different software companies have put forward their own coding standard, so the software is not compatible too, which also brings great inconvenience to users. All these reasons make the Uyghur information processing technology encountered countercurrent[12]. The Unicode standard should be widely publicized so as to realize the unification of Uyghur code.

The number of Uyghur websites are increasing, Uyghur language inputing method are used by more and more Uyghur citizens. With the popularity of these Uyghur websites and Uyghur applications, more Uyghur cityzens use the national language through the internet to express their emotions, attitudes, opinions and requirements. These network opinions can quickly reflect the public opinion situation and development trend of XinJiang. Therefore, the research on Uyghur internet hot topic and Uyghur word classification become a strong need. However, there has not been a system that is able to moniter Uyghur internet information.

What we can do next is to develop a hot topic tracking analysis system based on the topic detection and hot topic computation that can discover hot topic and realize the text classification. Firstly, we need to read a lot of literature to make a detailed analysis subject identification model, and combine with the characteristics of Uyghur web hot topic analysis, formulate a system framework, and set up the Uyghur hot topic detection and word type classification model; Secondly, we'll do a lot of research to unify the Uyghur coding, Uyghur smart segmentation and algorithms of Uyghur entity recognition, so that we can get the classification which can be used in Uyghur hot topic detection.

There are many aspects of technology about Uyghur information processing still in the preliminary stage, such as word segmentation, weaving search engine. These technology has got some achievements, but still no word segmentation standard of Uighur, word segmentation technology still needs further mature. Uyghur word segmentation is an important link of Uyghur information processing, is one of the foundation treatment Machine Translation, speech recognition, topic detection, intelligent retrieval, The problem of development of Uighur word segmentation technology will become a major obstacle of Uyghur information processing technology development, it is difficult to be solved. Although the Uigur search engine has begun to appear, but the results are not accurate enough, the search scope is not big enough. The domestic and foreign each big search engine, such as Baidu Google did not support the Uighur information search, this shows that the Uyghur information search engine technology has a long way to go. In addition, weaving information processing and Uyghur handwriting input recognition, popularity of Uighur software and many other areas where development is needed.

Conclusion

There is an important significance of Uyghur information processing technology, involves

interdisciplinary knowledge of computer science, information science, acoustics such as a large number of disciplines. This paper is devoted to the analysis of the process of Uyghur information by computer, is the study of the history of Uyghur information processing technology, research status, problems still exist, and the future prospect of Uyghur information processing technology. I hope this paper can have the inspiration to colleagues, but also hope to be able to communicate more learning, improving the technology better.

Acknowledgement

This work is supported by “the National Natural Science Foundation of China (No.61309012)”.

References

[1] WeiNiJiang·MuShaLa，AiErKen·YiMiEr. Online Process and Realization of Uygur Unicode[J] JOURNAL OF XINJIANG UNIVERSITY(NATURAL SCIENCE EDITION . 2004, 21(3)：331-335.

[2] YaSenjiang， Abuduaini. Uyghur Information Processing on Office 2000 [J]JOURNAL OF XINJIANG EMPLOYEE UNIVERSITY 2001, 9(2):56-58

[3] LuYouFei，ZhangWei，ZhangYan，MiaoCheng，LiChun. Study and Realizaton of Pivotal Technology in The Design of Uyghur Version Office[J]. JOURNAL OF CHINESE INFORMATION PROCESSING 2007, 21(2)：112-116.

[4] The Commerce Department of Xinjiang the Xinjiang Uygur Autonomous Region. 2014 electronic commerce development report. http://www. xjftec. gov. cn/tzgg/201411/t20141118_123516. html

[5] YANG Xiao-lan. QIAN Cheng. ZHAO Hai-ting. Explore and Analyse Web Text Processing Tecknology. [J]. COMPUTER KNOWLEDGE AND TECHNOLOGY . 2010, 06(29)：8247-8249

[6] HUANG Chang-ning. ZHAO Hai. Chinese Word Segmentation:A Decade Review. [J]. JOURNAL OF CHINESE INFORMATION PROCESSING 2007, 21(3):8-19.

[7] XUE Hua-jian, DONG Xing-hua, WANG Lei, TURGHUN · Osman，JIANG Tong-hai. Unsupervised Uyghur Word Segmentation Method Based on Affix Corpus[J]. Computer Engineering and Design. 2011, 32(9)：3191-3194.

[8] Gulila Adongbieke，Mijit Ablimit. Research on Uighur Word Segmentation[J]. JOURNAL OF CHINESE INFORMATION PROCESSING. 2004, 18(6)：61-65.

[9] CHENG Li-zhen，Kamil·Moydin. Research 0n the Uiaur Web Information RetrIevaI System[J]. COMPUTER KNOWLEDGE AND TECHNOLOGY。2005(12)：5-9.

[10]LI Yan-jiao， JIANG Tong-hai. Uyghur text classification model based on improved weighted Bayes. [J]. Computer Engineering and Design. 0 2012, 33(12)：4726-4730.

[11]Aisaitijiang Aibaidula. The Design of Uighur Search Engine and It's Carrying—Out[J]. OURNAL OF XINJIANG EDUCATION INSTITUTE. 2004, 20(4)：102-106.

[12]AbDulLa. The specific format of the database of Uyghur information processing and retrieval based on. [D] Shanghai. Tongji University School of software. 2007:1-53.