International Journal of Database Theory and Application
Vol.9, No.5 (2016), pp.45-54
http://dx.doi.org/10.14257/ijdta.2016.9.5.05

Research on Modern Uyghur Common Word Extraction

Azragul [1,2,3], Alim Murat[1,2,3] and Li xiao[1,2]

[1]The Xinjiang Technical Institute of Physics & Chemistry. CAS, Xinjiang Key Laboratory of Minority Speech and Language Information Processing, Urumqi, Xinjiang, 830011, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]School of Computer Science & Technology Xinjiang Normal University, Urumqi, Xinjiang, 830054, China)
Azragul2010@126.com

Abstract

The key techniques and methods for the construction of modern Uyghur language (MUL) corpus are presented. The techniques and methods included MUL corpus, MUL corpus pre-processing, MUL corpus statistics, MUL stemming and MUL data analysis; on the basis of related works we then developed an enhanced modern Uyghur Common Words (UCW)-glossary. We conducted basic inspections upon the words from two perspectives namely the usage frequency and distribution. Upon developing enhanced MUCW glossary we considered the number of word types, word frequency, word length, and the number of texts used as major factors.

Keywords*: Modern Uyghur language; Corpus; Common words lexicon; Quantitative analysis*

1. Introduction

Modern Uyghur Common Word Extraction is a basic project in the areas of Minority language information processing is necessary. Lacking of Uyghur Common Word (UCW)-glossary, is a major factor impacting on the quality of information processing tasks in Uyghur, in which computational linguistics and lexicology. Setting out a UCW glossary, with such representativeness, viability and authority, boosts the leap-forward development of Altaic Minority language family in Xinjiang, like Uyghur, Kazak, Kirghiz, *etc.*

The paper carried out a main research on corpuses origin and domain, In order to ensure those corpuses' reliability, typicality and authority.

By means of existing corpus resources, we built Modern Uyghur Corpus, which comprises publications, including textbooks and teaching materials, broadcast and cyberspace in a systematic and sustainable way, by collation and post-process. This corpus has different sources, wider range and appropriate control in various domains than the previous one.

The paper improved and completed the method and key technique for Modern Uyghur Corpus construction in further, and incorporated personal name recognition and automatic data analysis. Besides, conducting an initial investigation on word usage and word text-occurrences, moreover, in the light of word frequency and number of intext-occurrences of word, Modern UCW candidate list is presented.

This research not only provides a solid foundation for the task in areas of Minority language processing, such as Uyghur or it can offer services to language standardization, textbook design, literature teaching in primary and middle school, anti-illiteracy, bilingual education and dictionary compilation of Altaic language families.

ISSN: 2005-4270 IJDTA

2. Uyghur Language Corpus Constructions

In this section, on the basis of currently available corpus date as a selecting reference for different media, UCW candidate list is purposed while certifying its typicality and authority.

This corpus consists of four main media, such as the print media (refers to the literacy works and the classic masterpieces), the teaching media (refer to the formal publication related to many domain, like Science, Culture, Economy, Industry by Xinjiang Education Press, Xinjiang Science Press, People's Publishing House of Xinjiang and Fine Arts Publishing House),Broadcast (refers to the transcript of a 30-minutes Xinjiang news broadcast of daily program of Xinjiang TV) and the Web media (refers to the Cyber language from dozens of relatively formal website). Those four generally express majority part of Uyghur concerning with politics, economy and social life.

This corpora is provided by the Minority Language branch center for Uyghur language research base of National Language Resource Monitoring, and the key lab of Network Security and Public Opinion Analysis of Xinjiang Normal University.

2.1 The Print Media

In the print media, literacy works by national publishing house is considered as prime research object, and capacity is 188MB, 20.81% of total corpus.

2.2 The Teaching Media

In the teaching media, educational and instructive formal publication is considered as main research object, and capacity is 173MB, 24.67% of total corpus.

2.3 The Broadcast Media

In the broadcast media, the transcript of a daily 30-minutes Xinjiang news broadcast and a daily 30-minutes national news broadcast are used. The time-line of corpus collection spans from January, 2010 to December, 2012 for 1080days. And it including various news sites taken place in all different domain, where international, national, xinjiang, culture, education, economy, health, and social life, and capacity is 171.2MB, 24.42% of total corpus.

2.4 The Web Media

In the web media, websites, such as Xinjiang government official site, Kunlun site and Tianshan site, *etc.* considered as a major websites object. Corpus data, that gathered since April, 2006 to December, 2012, composed various domain, in which international, national, xinjiang, culture, education, economy, health, and social life, and capacity is 169MB, 24.10% of total corpus.

3. Modern Uyghur Common Word Extraction

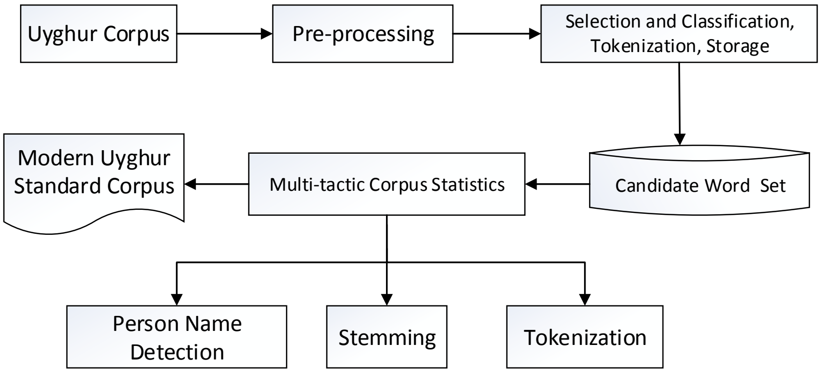
Uyghur Corpus
Pre-processing
Selection and Classification, Tokenization, Storage
Modern Uyghur Standard Corpus
Multi-tactic Corpus Statistics
Candidate Word Set
Person Name Detection
Stemming
Tokenization

Figure 1. Standardized Corpus Based Modern Uyghur Multi-Tactic Statistical Model

(1)Modern Uyghur Corpus pre-processing: selection, pre-processing and Textile generation.

(2)Modern Uyghur corpus statistics

①Statistics on corpus including word token, word frequency, word species, word length and word intext-occurrence

②Person name detection: according to the characteristics of Uyghur name, Chinese person name and foreign name, optimize person name detection. Determining the detection rules, solve the whitespace separation between First and last name in Chinese and improve accuracy of Chinese and Foreign name detection.

(3)Modern Uyghur word Stemming

Stemming: dictionary lookup and Human-computer interaction combined method.in terms of stemming, a newly appeared stem is discovered by making use of Modern Uyghur stem dictionary and added to the machine dictionary, in this way, systems machine learning ability is enhanced. Word formation in Uyghur shown in Figure 2.

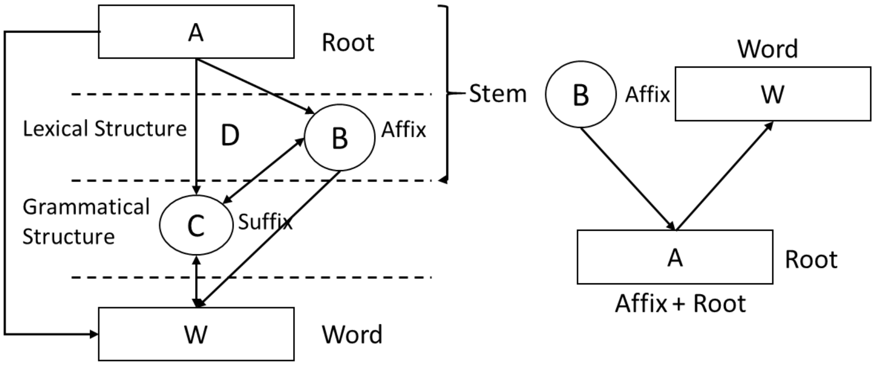

Figure 2. Modern Uyghur Lexical Structure Model

In Figure 2, the A, B, C, D and the W indicate the root, affix, suffix, stem and the word respectively.

(1)Data Analysis on Modern Uyghur

In this part, we make a calculation and analysis on the common word, sub common word, partial common word, stand-alone word and word species distribution and coverage.

Number of word token: a summation of each unique word's token count in the whole corpus. It is a specific number, which intuitively expresses a typical word's real and original usage in corpus.

$$A_i = \sum_{k=1}^{i} n_k$$

In equation, Ai is the summation of object i's token count, and ni is the number of occurrences in corpus.

$$\mathrm{Bi}= \sum_{k=1}^{i} n_k / N \times 100\%$$

Word frequency: the percentage of each word's token count in a total number of the whole corpus, that:

In equation, Bi indicates the frequency of object i, ni is object I's occurrence, N is o total occurrence of object in the whole corpus.

Generally, the higher frequency of word shows the higher degree of common word becomes. Intuitively, in this case, this is the most effective statistical method.

Accumulative word frequency: the percentage of frequencies of each word and the previous one in the whole corpus, while all words are listed in order from low-frequency to high-frequency.

$$F_i = \sum_{i=1} n_i / N \times 100\%$$

In equation, Fi indicates the coverage of object i, ni is object i's occurrences is a total occurrence of object in the whole corpus. Note that the key role of accumulation frequency is to observe the each word position in the whole word list. While words are listed from high-frequency to low-frequency.

(2) Quantitative analysis on word domain generality

Word domain generality is to measure the degree of generality of a word in various language domain, that is to say, a quantified measure of a word's generality degree.in terms of calculation, it investigates word frequencies, while if a word is evenly distributed in multiple document and various domain.

In this paper, we adopt the modified domain generality algorithm as follows:

①compute domain oriented word occurrence Fx:

Fx indicates the total frequency of the k-th word in a domain oriented corpus.

②compute text usage of the k-th word UIX:

Compute the word text-usage using A.Julland formula:

Word's text usage:

In text usage, it describes relative occurrence of k-th word in i-th domain class, and indicates average relative occurrence of k-th word in all domain class, moreover, n is a total document of corpus. It represents scatter coefficient of the k-th word, and frequency of the k-th word.

③compute domain generality of the k-th word:

Compute the uniformity of a word distribution in all domain class employing distributional uniformity, computational formulas as follows:

Distributional uniformity:

Given SMR and Mean definition respectively, as follows:

Domain generality of k-th word:

In above equation, n is domain oriented class number, Fki is the number of k-th word occurrence in i-th domain class, Ulk indicates the text usage of k-th word, DCk represents the distributional uniformity of k-th word's domain class.

(3)Quantitative analysis on word time generality

Word time generality is quantified indicator measuring the observation word generality in investigation period. It requires whether observation word is used stable or not in investigation period, that is to say, monthly distributional uniformity of a word's occurrence.

Time generality computing process as follows:

①Compute monthly frequency of word Fk：

Fk represents a monthly total occurrences of the k-th word in the corpus.

②compute time generality of the k-th word Tk:

In terms of computation, distributional uniformity is adopted to compute the monthly distributional uniformity of a word in investigation period, computational formulas as follows:

(3)the k-th word time generality:

In above equation, n indicate number of month in investigation period, which require quantity of the corpus selected by monthly; Fki is word frequency in the i-th month.

(4)Quantitative analysis on word generality

Word generality is purposed considering word domain oriented usage and time stability as a whole, not including the impact from the regional generality and word generality. The regional generality should be incorporated into, later when considering large regional scale corpus.

Word generality computation as follows:

In word generality, the Tk, Uk and Ok represent time generality of the k-th word, the domain generality of the k-th word and word generality. The higher generality yields the better characteristic feature that can show commonly usage and stability of the k-th word in investigation period.

4. Modern Uyghur Common Word Lexicon Construction

We initiate a basic investigation on word and take both word usage and word distribution into account. UCW candidate list consider number of word species, number of word token, word frequency, number of word intext occurrences and word length as its basic fact of Uyghur language.

On that basis, we extract high frequency wordlist from all corpus in different media. At the first, we select the same words for all corpus in four different media, as a UCW candidate list; Then select the same words for any three corpus, as a sub UCW candidate list; then again, select the same words for any two corpus, as a partial UCW candidate list; in the end, select stand-alone word for each corpus, as a partial UCW candidate list.

5. Experimental Data

5.1 Bais data

Four big media, in which print media, broadcast media, internet media and textbook and teaching material, cover the corpus of this study and that comprise 96,025 text file and 43,529,435 word token. The basis of Modern Uyghur corpus selection and construction described in Section 3.

The corpus we built in this paper composes of print media, textbook and teaching material, broadcast media and internet media. On the whole, this corpus involve every parts of Uyghur's life with respect to the politics, economy and social activities. Corpus details shown in Table 5-1.

Table 5-1. The Distribution of Total Corpus

Sources	Print Media	Teaching Media	Broadcast	Web	Totals
#Word-token	11 879 662	12 195 468	10 587 381	8 866 924	43 529 435
#Word species	350 760	273 230	216 021	323 660	703 669
#Word-stem species	106 386	91 892	68 053	78 333	147 054

5.2 Comparisons between the Common Word and Word Stem

In order to ensure typicality and authority of the UCW candidate list, either a Uyghur word or Uyghur word stem can be regard as UCW. We conduct a comparative analysis on the corpus, according to the characteristics of Uyghur and details of the four media.

(1)Basic Statistics of Modern Uyghur Word

Make a comparison between the four wordlists of the four media including print media, textbooks and teaching materials, broadcast and internet media, by extracting the same words for the four media. It is composed of 62,330 UCW.

Table 5-2. UCW Basic Statistics

Item	#Word species	Percentage (%)	#Word token	Percentage (%)
UCW	62 330	0.18	33 834 388	77.73

From Table 5-2, the given fact that there is only 62,330 same word for the four media covering 77.73% of the total corpus, show relatively lower coverage to the corpus, however, it doesn't undertake certain role of the candidate words of the UCW.

(2) Basic Statistics of the four media's word stem

Similarly, make a comparison between the four word-stem lists of the four media, by extracting the same word-stems. It is composed of 36,488 UCW candidate word stem.

Table 5-3. UCWS Basic Statistics

Item	#Word-stem species	Percentage (%)	#Word-stem token	Percentage (%)
UCWS	36 488	24.85	41 452 953	95.23

From Table 5-3, the given fact that there is 36,448 same word-stem for the four media covering 95.23% of the total corpus, show a coverage exponential close to the whole corpus, however, it can take certain role of the candidate words of the UCW.

5.3 Basic Statistics on Uyghur High-Frequency Word and Word-Stem

High-frequency word is refer to as all words whose total word frequency coverage in corpus reached 90%. So by this definition, we extract high-frequency word and word-stem from each type of corpus media, covering 90% of the total. The details shown in Table 5-4.

Table 5-4. The Distribution of High-Frequency Word and Word-Stem

Item	#Word token	Percentage (%)	#Word species	Percentage (%)	#Word-stem	Percentage (%)
Print Media	10 691 698	24.56	43 901	6.24	12 224	8.31
Teaching Media	10 975 954	25.22	24 233	3.44	9 561	6.5
Broadcast	9 528 624	21.89	12 794	1.82	4 595	3.12
Web Media	7 980 238	18.33	29 398	4.18	8 165	5.55
Totals	43 529 435	90.00	703 669	15.68	147 054	23.48

In this paper, we divide UCW candidate words into four classes like common word, sub-common word, semi-sub common word and stand-alone word, on the basis of the fact that word-stem can take a role of UCW candidate words. In computing, teaching media, print media, web media and broadcast are described as A,B,C and D respectively. The same words for all media is defined as UCW candidate words; the same words for any three media like (ABC, ABD, ACD, BCD) and any two media like (AB, AC, AD, BC, BD, CD) sub-UCW candidate list; Most of the words only come up in one media corpus like (A, B, C, D) is defined as stand-alone words. After that we extract the UCW candidate list and sub-UCW candidate list by comparison between high-frequency word-stem in the four media corpus.

Table 5-5. UCW Candidate Words and Stand-Alone Words

Item	#Word species	Percentage (%)	#Word token	Percentage (%)
UCW candidate word	3 186	19.9	30 468 709	77.77
Sub-UCW candidate word	5 889	36.79	6 861 820	17.52
Stand-alone Word	6 934	43.31	1 845 738	4.71
Totals	16 009	100	39 176 267	100.00

From Table 5-5, the given UCW and stand-alone words distribution. Due to the fact that combination of the UCW-candidate words and sub-UCW candidate list cover 90.22% of total high-frequency word in corpus, it shows that the extracted UCW candidate list is feasible in observed corpus.

5.4 Modern UCW-Glossary

In Table 5-6 below, there are 22 high-frequency words gained more than 10 thousand times occurrences in corpus.

Table 5-6. Sample of 22 High-Frequency Words Gained More than 10thousand Times Occurrences

Word	Chinese Translation	#Word Token	inText-occurrences	Length	Distribution	English Translation
ئۇ	他、她、它	563 989	42 862	1	A,B,C,D	He, She, it
ۋە	和	542 404	63 428	2	A,B,C,D	And
بۇ	这、这个	508 802	64 100	2	A,B,C,D	This
بىلەن	与	470 366	60 264	5	A,B,C,D	With
بىر	一	436 498	54 179	3	A,B,C,D	One
قىلىش	做（将来）	191 621	38 740	5	A,B,C,D	Will do
بولۇپ	做（过去）	185 521	38 089	5	A,B,C,D	Did
قىلىپ	做（过去）	149 610	35 071	5	A,B,C,D	Did
بولغان	做过	145 446	33 361	6	A,B,C,D	Have done
دەپ	说	137 666	19 444	3	A,B,C,D	Say
شىنجاڭ	新疆	121 936	28 032	6	A,B,C,D	Xinjiang
كېيىن	以后	121 102	29 219	5	A,B,C,D	Later
دۆلەت	国家	119 765	30 888	5	A,B,C,D	Nation
ئىككى	二	119 575	27 166	4	A,B,C,D	Two
كېرەك	必须	119 571	18 168	5	A,B,C,D	Must
بولىدۇ	行、可以	119 300	17 380	6	A,B,C,D	Can, May
ئۈچۈن	为	115 657	27 821	4	A,B,C,D	For
ئادەم	人	114 039	20 875	4	A,B,C,D	Man
خەلق	人民	111 886	24 937	4	A,B,C,D	People
جۇڭگو	中国	105 568	26 138	5	A,B,C,D	China
ئاپتونوم	自治	102 111	23 752	7	A,B,C,D	Autonomy
شۇ	是、就是	101 740	21 616	2	A,B,C,D	That is

6. Conclusion

In this work, we built Modern Uyghur language corpus comprising Print media, Teaching media, Broadcast and Web media, throughout selecting a large-scale authentic language materials with help of the Uyghur language research base, and that of 43,529,435 word tokens. At current stage, utilizing these resources in a proper and effective way is of great significance to deepen and extend the language resources monitoring task for Uyghur, simultaneously, important reflection and beneficial trail with respect to the language life, language teaching, language engineering and dictionary compilation, *etc.* that computational linguistics serves. Where by, the four media's attention to the changes in people social life by means of the changes in frequency and relative changes in frequency-sequence ranking, give us brief information relating to the social life and current affairs throughout the usage of these words.

Acknowledgments

This work in this paper supported by National Social Science Foundation of China (No.14AZD11); National Nature Science Foundation of China (No. 61262066, 61132009); Natural Science Foundation of Xinjiang Uyghur Autonomous Region (No. 2014211A045); Social Science Foundation of Xinjiang Uyghur Autonomous Region (No. 14CYY093); General Social Science project of Ministry of Education (No. 14YJC740001).

References

[1] P. Yuequn, “Research on temporal Information Recognition and Normalization”, Harbin Institute of Technology, **(2006)**.
[2] W. Yun, “Chinese Temporal Information Extraction in Financial Field”, Tsinghua University, **(2004)**.
[3] W. Tong, “Research on Chinese Time Expression Recognition”, Fudan University, **(2010)**.
[4] L. Junchan, T. Hongye and W. Fenge, “Recognition of Temporal Expression and their Types in Chinese”, Computer Scince, vol. 39, no. 11, **(2012)**, pp. 191-194.
[5] L. Jie, “The express of Past Tense Contrast between Uyghur and Chinese”, Xinjiang University, **(2007)**.
[6] H. Tomur, “Modern Uyghur Grammar”, Ethnic Publishing House, **(1987)**, pp. 325-370.
[7] Y. Lei and C. Junyi, “Study on Uyghur Time Adverb”, Language Research, no. 10, **(2012)**, pp. 155-156.
[8] W. Fuling, “Comparative Study on Temporal Expression Rule between Uyghur and Chinese”, Language and Translation, no. 4, **(1994)**, pp. 27-31.
[9] A. Kadir, “Comparative analysis on Temporal Noun between Uyghur and Chinese”, Journal of Changchun University, vol. 23, no. 7, **(2013)**, pp. 836-838.
[10] H. Ruifang, Q. Bing, L. Ting, P. Yueqin and L. Sheng, “Recognizing the Extent of Chinese Time Expression Based on the Dependency Parsing and Error-Driven Learning”, Journal of Chinese Information Processing, vol. 21, **(2007)**, pp. 36-40.
[11] Z. Guorong, “Research on Temporal Expression of Chinese News”, Shanxi University, **(2006)**.
[12] W. Tong, Z. Yaqian, H. Xuanjing and W. Lide, “Chinese Time Expression Recognition Based on Automatically Generated Basic Time Unit Rules”, Journal of Chinese Information Processing, vol. 24, no. 4, **(2010)**, pp. 4-10.

Authors

Azragul, was born on October 12, 1987, in Xinjiang, China. She is currently reading a doctorate in Computer Applications Technology at China's Academy of Sciences. Her main research fields are computational linguistics and natural language processing. Email:Azragul2010@126.com

Alim Murat, was born on October 15, 1988, in Xinjiang, China. He is currently reading a doctorate in Computer Applications Technology at China's Academy of Sciences. Her main research fields are computational linguistics and natural language processing.

Li Xiao, corresponding author was born on September 6, 1957, in Xinjiang, China. He is an researcher’r in The Xinjiang Technical Institute of Physics & Chemistry. CAS. His main research fields are Multilingual information processing.