International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 18, No. 1, 2025

Research on the Selection Method and Data Set Construction of Yi Language Stop Words

Junping Huang

Fangchenggang City, Guangxi Zhuang Autonomous Region, China

Abstract: This paper aimed to complete the research on the selection method and data set construction of the Yi language stop words which has positive significance for information retrieval, text classification, sentiment analysis and online public opinion analysis. On the basis of word segmentation, data cleaning and word frequency statistics of Yi language text, used statistical methods such as Document Frequency, Term Frequency, and Entropy, analyzed and selected the top 100 high-frequency words that were ranked by these three measures, through the intersection of three sets of high-frequency words, 89 words were identified as potential stop words in the Yi language, and according to the characteristics of stop words, 20 function words were selected as stop words. Based on the methods of selecting stop words in Chinese, Mongolian, Tibetan, Uyghur, etc., completed the construction of the Yi language stop words dataset.

Keywords: Yi Language, Stop Words, Selection Method, Data Set, Construction.

1. Introduction

Stop words refer to a type of word that appears frequently in a specific text and whose semantic information has little impact on Natural Language Processing. Stop words are pivotal in the Natural Language Processing of the Yi language, and can be used in tasks such as information retrieval, text classification, sentiment analysis, and online public opinion analysis. In Natural Language Processing, after word segmentation, eliminate stop words from the text before further processing will significantly reduce the text's dimensionality, enhance the computational efficiency of the computer, and consequently improve the speed, quality, and accuracy of text processing. This study explores the method of selecting stop words in Yi language from a certain amount of Yi language texts that have completed word segmentation, data cleaning, and word frequency statistics, and constructs a dataset of stop words in Yi language.

2. Related Work

Abroad, the study of stop words initially arose from research in information retrieval. Luhn [1], [2] found in his research on information retrieval that some words appear frequently, but these words seriously affect the retrieval performance, he proposed using noise to represent such words. Wilbur et al. [3] identified stop words in a collection using automated statistical testing. Ho [4] proposed a stop words localization and recognition strategy in adaptive text recognition. Khabibull et al. [5] used DF and TF statistical algorithms to construct a stop words dataset for Karakalpak language.

In the study of Chinese stop words, Hao et al. [6] proposed to use Chi-square test to complete text classification and automatic extraction of stop words lists. Zou et al. [7] proposed a stop words selection method based on statistical and information theory models. Jiang Zhaozhong [8] studied Chinese automatic word segmentation based on context and stop words drive, and pointed out that Chinese stop words refers to words with high frequency and no great retrieval significance. Xiong Wenxin et al. [9] studied the stop word filtering of information retrieval users' query statements. Zhou Qinqiang et al. [10] used the program flow control to eliminate the single independent words, English characters, numbers, a series of mathematical symbols and the Chinese words containing these symbols, so that the pure Chinese words with more than two characters became the feature items representing text information. Yang et al. [11] conducted a comparative study on feature selection methods in the statistical learning of text classification, and carried out dimensionality reduction processing on text.

Research on stop words in Chinese minority languages, Gong Zheng et al. [12], [13] obtained a list of Mongolian stop words through statistical methods, in which there were often some entity nouns that were closely related to the subject but affected the retrieval accuracy, and used the Joint Entropy algorithm to preliminarily determine the Mongolian stop words, after removed the entity nouns and homonyms from the initial determination of Mongolian stop words, the list of Mongolian stop words was determined by comparing the parts of speech of English stop words and Mongolian stop words. Zhu Jie et al. [14] analyzed the selection of stop words in Tibetan through methods such as Term Frequency, Document Frequency, Entropy, etc., and proposed a Tibetan stop words selection method that combined Tibetan function words, special verbs, and automatic processing methods. SAIMAITI et al. [15] analyzed the characteristics of Uyghur stop words, used the methods of Document Frequency, Term Frequency and Entropy to conduct statistics on a large number of corpus, and analyzed the part of speech distribution of candidate stop words.

3. Stop Words Selection Method

Used statistical methods such as Document Frequency (DF), Term Frequency (TF), and Entropy (EN) to automatically select stop words in Yi language. All three methods apply statistical methods to assess the importance of a word to a text set, and the importance of a word increases in proportion to the number of times it appears in the text set.

3.1. Document Frequency Method

The Document Frequency method counts how many texts

a word appears in a text set. The calculation formula is as follows:

$$DF(x_i)=\sum_{j=1}^{m}\omega(x_i,y_j) \tag{1}$$

$\omega(x_i,y_j)$ indicates whether the word x_i appears in the document y_j. If the word x_i has appeared, the value of $\omega(x_i,x_j)$ is taken as 1, otherwise it taken as 0, and m is the total amount of texts.

3.2. Term Frequency Method

Term Frequency refers to the number of times a word appears in a given document, abbreviated as word frequency. In order to prevent excessive word frequency in long texts, which may affect the calculation results, it is usually necessary to normalize the word frequency, the specific method is to divide the word frequency by the total number of words in the text. The formula is as follows:

$$\omega_{ij}=\frac{x_{ij}}{\sum_{i=1}^{n}x_{ij}} \tag{2}$$

x_{ij} represents the number of times word x_i appears in text j, the denominator is the sum of times each word appears in text j, and n is the total number of words in text j. The weight of a word in all texts is calculated by the following formula:

$$TF(x_i)=\sum_{1\le j\le m}\omega_{ij} \tag{3}$$

$TF(x_i)$ represents the weight of the word x_i in all texts, ω_{ij} represents the weight of the word x_i in text j, and m is the number of total texts.

3.3. Entropy Method

Entropy is a concept introduced by Claude Shannon in information theory to quantify the uncertainty of information, the greater the uncertainty, the greater the entropy, and the lower the entropy. The formula for calculating the entropy of word x_i in m texts is as follows:

$$EN(x_i)=1+\frac{1}{\ln m}\sum_{j=1}^{m}p(x_{ij})\times\ln p(x_{ij}) \tag{4}$$

$$p(x_{ij})=\frac{f_j(x_i)}{\sum_{j=1}^{m}f_j(x_i)} \tag{5}$$

$f_j(x_i)$ is the word frequency of word x_i in text j, and m is the number of total texts.The entropy calculation result of a certain word is processed in ascending order, that is, the higher the word frequency, the smaller the entropy value.

4. Experimental Design

4.1. Data Sources

The data came from Yi language newspapers and books, which covered six categories such as news, politics, agriculture, culture, literature and tourism. Each category contained 10 texts, a total of 60 texts, each text contained 3000-7000 words Yi language characters, , with a total word count of 574158.

4.2. Data Pre-processing

The preprocessing of Yi language data included processes such as word segmentation, data cleaning, and word frequency statistics.

(1) The Yi language word segmentation was carried out using the Bi-directional Maximum Matching (Bi-MM) method based on the Yi language dictionary which contains 18,356 Yi language words. This method compares the results of the Forward Maximum Matching (FMM) method and the Backward Maximum Matching (BMM) method, and takes the result of the least number of words as the final result. There were certain segmentation errors in the o-btained segmentation results, which needed to be corrected manually. Partial word segmentati-on results are as follows:

/ꆏ㐅/ꑭ㠯/ꆏ㕘/ꎾ/ꉬ㑇/ꑍ㖞/ꃅ㖞/ꀋ/“/ꆏ㕘/ꎾ/ꑍ㖞/：/ꁦ㐅/ꁨꑍ/ꆏ㐅/ꑭ㠯/ꑍ㖞/ꎾ/”，/ꀋ㖞/ꂷ㚔ꋍ/ꑍ㖞/ꑍ㖞/ꑍ/ꑍ/ꑍ/，/25ꀕ/ꓐ㖞/ꑍ㖞/ꑍ/。/

(2) Data cleaning referred to the process of cleaning segmented text, including removing punctuation marks, numbers, and English letters from the text.

(3) Word frequency statistics was the process of counting the occurrence of Yi language vocabulary in text that had undergone word segmentation and data cleaning. Because of the low efficiency of manual statistics and the problem of missing statistics, thus a computer program was written to process the text, and every non-Yi language character was replaced with a newline number "\n", so that each Yi language word in the text was stored in the txt file by line, the blank line in the text was removed, and counted the frequency of each Yi language vocabulary appearing in the text, and outputed in order of size. The partial output results are as follows:

The number of Yi language words contain in this text is: 1129. Yi language vocabulary [Quantities, Yi language]:[226,'ꑍ'][154,'ꑭ'][123,'ꑍ'][120,'ꑌ'][100,'ꑌ'][54,'ꑀꑣ'][33,'ꑀ'][29,'ꑩꑾ'][28,'ꓐ']

After lexical statistics of 60 files containing 574158 Yi language words, 5777 Yi language words were obtained, with a total word frequency of 175628, the word frequency of the top 100 words was 60486, accounting for about 34.44% of the total word frequency, and the word frequency of 1 was 1541, accounting for about 0.88% of the total word frequency. It can be seen that the top 100 Yi language words accounted for a large proportion of the total word frequency, so the top 100 Yi language words were selected as high-frequency words.

4.3. Experimental Analysis

Three statistical methods, DF, TF and EN, were used to process the text, and the DF and TF calculation results were sorted by descending order. Because the entropy of EN calculation results was smaller, the word frequency was

higher, so the EN calculation results were sorted by ascending method. Each method selected the top 100 high-frequency words as statistical results.

(1) DF top 100 high-frequency words: [illegible]

(2) TF Top 100 high-frequency words: [illegible]

(3) EN Top 100 high-frequency words: [illegible]

Performed high-frequency word intersection analysis on the statistical results of DF, TF, and EN, and obtained the following table:

Table 1. The intersection analysis of DF, TF and EN high-frequency words

Set intersection	Quantity	Proportion(%)
DF ∩ TF	91	91%
DF ∩ EN	94	94%
TF ∩ EN	93	93%
DF ∩ TF∩EN	89	89%

As can be seen from Table 1, the high-frequency words selected by three statistical methods, DF, TF and EN, have little difference. The overlap degree between DF and TF is 91%, the overlap degree between DF and EN is 94%, the overlap degree between TF and EN is 93%, and the overlap degree between DF, TF and EN is 89%. It can be seen that the overlap degree of high-frequency words selected by the three statistical methods is high, and the difference is that the order of high-frequency words selected is different.

According to the characteristics of high frequency and no practical meaning of stop words, perform part of speech analysis on the intersection of high-frequency words (89 words) selected by DF, TF, and EN methods, and finds that some function words have high frequency, and the function words have no meaning when used alone. Therefore, Yi language function words can be listed as stop words. At the same time, it also proves that using statistical methods such as DF, TF, EN to select stop words in Yi language is effective. Among the 89 high-frequency words that can be used as function words, there are 20 words such as [illegible], etc., accounting for approximately 22.5% of the 89 high-frequency vocabulary words in Yi language, with a word frequency of 26405 times, accounting for approximately 15% of the total word frequency. [illegible] can be used as prepositions; [illegible] can be used as conjunctions; [illegible] can be used as auxiliary verbs; Credit, [illegible] can be used as adverbs. Some Yi language words have multiple parts of speech, for example: [illegible] has prepositions, pronouns and verbs; [illegible] has prepositions, verbs, quantifiers, adjectives and conjunctions. In this study, Yi language words with multiple parts of speech are temporarily treated as function words.

It can be seen from the experiment that a large part of Yi language function words appear less frequently, so it is difficult to select them by automatic selection method, so all Yi language function words can be classified as stop words. According to the selection of stops words in Chinese [9], Mongolian [12], Tibetan [14], Uyghur [15], etc., pronouns in Yi language can be listed as stop words in Yi language, and numbers and punctuation marks can also be stop words. In addition, for different fields of Yi language research, statistical methods such as DF, TF, EN can be used to select other high-frequency vocabulary as stop words. According to the needs of research, stops words can also be nouns, verbs, adjectives and other content words.

5. Yi Language Stop Words Data Set

Through the collection, sorting, classification and summary of Yi language words in Yi language dictionaries, grammar books and textbooks, the data set of Yi language stop words is obtained as follows:

(1) Function words:

① Conjunctions (33): [illegible]

② Prepositions (17): [illegible]

③ Interjections (35): [illegible]

④ Auxiliary words (49): [illegible]

⑤ Adverbs (243): [illegible]

(2) Pronouns (65): [illegible]

[illegible]

(3) Symbols, numbers (23) :$? _ “ ”、。《》!，:； 0 1 2 3 4 5 6 7 8 9

6. Conclusion and Prospect

This paper proposed using statistical methods such as DF, TF, EN to selected stop words in Yi language, and 20 Yi language function words were selected as stop words according to the experimental results. Due to the low frequency of some Yi language function words, the automatic selection of stop words cannot be completely selected. Therefore, according to the characteristics of stop words, Yi language function words (prepositions, conjunctions, interjections, auxiliary words, adverbs) can be listed as Yi language stop words. Finally, Yi language stop words data set was constructed according to Chinese, Mongolian, Tibetan and Uyghur stop words selection methods.

In the future, there is still work to be done on the study of stop words in Yi language:

(1) Because some Yi language words have multiple parts of speech, which leads to inaccurate parts of speech recognition, it can strengthen the research on automatic parts of speech recognition technology of Yi language words, and build a more accurate data set of Yi language words.

(2) The Yi language function words and content words can be deeply explored in order to expand the Yi language stop words data set.

(3) On the basis of the study of stop words, the tasks of Yi language text information retrieval, text classification, emotion analysis and online public opinion analysis can be studied.

References

[1] Luhn H P .A Statistical Approach to Mechanized Encoding and Searching of Literary Inf-ormation[J].Ibm Journal of Research and Development, 1957, 1(4):309-317.DOI:10.1147/ rd.14.03 09.

[2] Luhn, H. P .The Automatic Creation of Literature Abstracts[J]. IBM Journal of Research and Development, 1958, 2(2):P.159-165.DOI:10.1147/rd.22.0159.

[3] Wilbur W J, Sirotkin K .The automatic identification of stop words[J]. 1992.DOI:10.1177/016555159201800106.

[4] Ho T K .Stop Word Location and Identification for Adaptive Text Recognition[J].International Journal on Document Analysis & Recognition, 2000, 3(1):16-26.DOI:10.1007/ PL00013551.

[5] Khabibulla M ,Shukurla B ,Jernej V .Dataset of Karakalpak language stop words[J].Data in Brief,2023,48109111-109111.

[6] Hao L , Hao L .Automatic Identification of Stop Words in Chinese Text Classification[J].IEEE Computer Society, 2008.DOI:10.1109/CSSE.2008.829.

[7] Zou F , Wang F L , Deng X ,et al.Automatic construction of Chinese stop word list[J].World Scientific and Engineering Academy and Society (WSEAS), 2006.

[8] Jiang Zhaozhong .Chinese words segmentation based on context and stopwords[D].Hefei University of Technology, 2010.

[9] Xiong Wenxin, Song Rou .Removal of Stop word in users' request for information retrieval [J].Computer Engineering, 2007,(06):195-197.

[10] Zhou Qinqiang, Sun Bingda, Wang Yi .Study on new pretreatment method for Chinese text classification system[J]. Application Research of Computers,2005,(02):85-86.

[11] Yang Y , Pedersen J O .A Comparative Study on Feature Selection in Text Categorization[J]. 1998.

[12] Zheng G , Gaowa G .The selection of Mongolian stop words[C]//IEEE International Conf-erence on Intelligent Computing & Intelligent Systems.IEEE, 2010.DOI:10.1109/ ICICISYS.2010.5658841.

[13] Gong Zheng, Guan Gaowa .A comparative study on between Mongolian stop words and English stop Words[J].Journal of Chinese Information Processing,2011,25(04):35-38.

[14] Zhu Jie, LI Tianrui .Research on Tibetan stop words selection and automatic processing method[J].Journal of Chinese Information Processing,2015,29(02):125-132.

[15] SAIMAITI Maimaitimin, ESMEL Abdureimu .Research on Uyghur stop words extraction method[J].Computer Engineering, 2019, 45(10) : 288-292,300.