Automated Stopwords Identification in Punjabi Documents

Rajeev Puri[1], Dr. R.P.S. Bedi[2], Dr. Vishal Goyal[3]

Research Scholar, Punjab Technical University, Kapurthala Road, Jalandhar.
Research Supervisor, Punjab Technical University, Kapurthala Road, Jalandhar
Assistant Professor, Dept of Comp. Sc, Punjabi University, Patiala.
rpuri@davjalandhar.com, bedirps2000@yahoo.com, Vishal.pup@gmail.com

Keywords: Punjabi Stop words list, statistical modeling, Borda count, Information Processing, Text classification

ABSTRACT

Many information retrieval tasks deal with the classification of huge amount of data before giving final results. The data being processed in IR tasks may or may not be useful for the researchers. There has to be some method to identify such data (called stop words) and remove it from data set before beginning with the IR task. This gives dual benefits – Reducing the overall vector space, thereby leading to performance improvements in terms of execution speed and the relevance of results. The purpose of this paper is to find a suitable, automated method for identification of stop words in Punjabi Text.

1. Introduction

In broader sense, stop words are the words which occur most frequently in a document and are of very little or no relevance to the text processing tasks. The Punjabi words such as ਦੇ, ਹੈ, ਦੀ, ਨੂੰ, ਸਿੰਘ, ਤੇ, ਵਿਚ, ਦਾ, ਨੇ, ਅਤੇ, ਇਸ, ਤੋਂ etc are a few most frequently occurring words in Punjabi text. A list of such words is known as Stop Words List. Such words account for the huge size of the corpus. These words should be removed in the preprocessing phase of the text classification process. This will not only decrease the vector space of the corpus, but will also speed up the calculations and increase the accuracy of the IR task. A stop word list can be manually constructed by analyzing the text. Since, the manual stop word construction is a time consuming process, which is further proportional to the size of the corpus under consideration, the automated methods should be preferred in place of the manual ones.

The researchers have suggested some statistical methods for automatically finding stop words in English text. A number of stop word lists based on these methods are available for English language documents [1]. These lists have also been adopted as standard stop word lists in may research works. Majority of these lists are prepared using the frequency analysis of words[2]. The constituents of stop word lists may also vary based on the domain of the corpus under consideration. A number of other stop word identification methods have been suggested for English language. Lo et all [3] have suggested a term based random sampling approach. Sinka and Corne [4] have suggested term entropy measure for finding stop words.

In comparison to the English language, not much extensive work has been done for finding stop words in Punjabi language. Gupta-Lehal [5] used frequency count for finding stop words in Punjabi text. The frequency count however cannot be taken as the true measure of stop word identification. In this paper, a revised statistical approach is suggested that along with frequency count, also takes care of distribution of stop words across the documents.

2. Construction of Stop words List in Punjabi

The stop words are the words that frequently occur in documents, but do not carry any significant information for Information Retrieval task. According to R.K. Blew [7], these words carry very poor index values. Users very rarely ask for documents with these terms. Moreover, these words make up a large fraction of the text of most documents. According to Francis and Kucera [8], the ten most frequently occurring words in English typically account for 20 to 30 percent of the tokens in a document.

In statistical method, a quick stop words list can be obtained by finding the most frequent words in the document. In another approach, the probability distribution of words over the documents can be analyzed and result can be used to predict the possibility of a word being a stop word. The two approaches can then be combined to find the aggregate stop words list. In subsequent sections of the paper, this methodology is discussed in details.

2.1 Experimental setup

For experimental setup, total 10000 news articles were taken from a Punjabi newspaper “Ajit”, with average 400 words per article. As a pre-processing phase for finding stop words in Punjabi, all the special characters, digits and punctuation marks were stripped from the document.

3. Classification of Stop words using Frequency analysis

After tokenization, a frequency count of words was prepared. The 50 most frequently occurring words are shown in Table 1.

Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency
ਹੈ	95798	ਨੇ	36286	ਉਸ	21857	ਆਪਣੇ	12216	ਜੋ	8330
ਦੇ	87408	ਹਨ	35156	ਤਾਂ	21750	ਕਰ	12200	ਕਰਕੇ	8066
ਦੀ	77420	ਨਾਲ	33626	ਇਹ	21270	ਪਰ	11961	ਕੋਈ	7949
ਨੂੰ	70581	ਤੋਂ	33436	ਨਹੀਂ	20685	ਗਿਆ	11835	ਰਹੇ	7642
ਵਿਚ	66022	ਸਿੰਘ	30618	ਇਕ	19486	ਦੀਆਂ	11424	ਜਾਂ	7573
ਦਾ	55207	ਕਿ	29302	ਚ	19042	ਕੀਤਾ	10611	ਕੀਤੀ	7569
ਤੇ	54526	ਕੇ	29238	ਹੋ	17434	ਨਾ	10500	ਜਾਂਦਾ	7318
ਅਤੇ	42503	ਹੀ	25980	ਉਹ	16067	ਜਾ	8912	ਵਾਲੇ	7301
ਇਸ	37599	ਲਈ	25177	ਉਨ੍ਹਾਂ	14859	ਸਨ	8800	ਵੱਲੋਂ	7275
ਵੀ	36361	ਸੀ	22757	ਕਰਨ	12870	ਜਿਸ	8550	ਇਨ੍ਹਾਂ	7163

Table 1: Most frequent Punjabi words along with their frequency in descending order.

Since the frequency count alone cannot be taken as the true classifier for the stop words, the distribution of the words in the documents also needs to be considered.

4. Statistical model based on distribution of words in documents

The mean and the variance can be used as indicators of distribution of words in the documents. The Table 2 shows the mean and variance of the top 50 words in descending order of their variance.

	Mean	Var		Mean	Var		Mean	Var		Mean	Var		Mean	Var
ਆ	28.87	626.06	ਜਾ	7.75	70.26	ਕੇ	5.49	28.45	ਜੀ	2.34	13.33	ਉਨ੍ਹਾਂ	1.57	7.65
ਚ	20.19	297.96	ਕਿ	7.44	64.62	ਹਾਂ	4.23	26.71	ਪਰ	2.80	12.83	ਕਰਨ	1.95	6.57
ਦੇ	15.40	199.25	ਨੂੰ	8.02	59.06	ਵੀ	4.71	24.80	ਫ਼ਿਲਮ	0.75	11.37	ਗੁਰੂ	0.58	6.44
ਦੀ	14.02	179.24	ਕਰ	7.88	55.34	ਇਸ	4.24	24.43	ਇਹ	2.35	11.13	ਹਰ	1.95	6.38
ਦਾ	13.19	169.73	ਹੋ	5.99	38.81	ਆਪ	3.12	18.45	ਇਕ	2.60	10.84	ਗਿਆ	1.63	5.76
ਹੈ	10.04	126.93	ਸੀ	4.52	36.99	ਤਾਂ	3.27	17.81	ਨਹੀਂ	2.21	10.58	ਸਾਹਿਬ	0.66	5.69
ਤੇ	11.76	116.67	ਸਿੰਘ	3.16	35.64	ਤੋਂ	3.71	17.65	ਲਈ	2.65	10.23	ਪਾਣੀ	0.57	5.14
ਨਾ	9.75	83.02	ਅਤੇ	4.29	34.70	ਉਸ	2.43	17.58	ਦੀਆਂ	1.96	9.60	ਸਨ	1.24	4.97
ਹੀ	8.04	80.40	ਨੇ	5.14	32.84	ਜਾਂ	2.68	17.20	ਉਹ	1.85	9.15	ਜੋ	1.51	4.62

ਵਿਚ	7.44	75.82	ਹਨ	4.07	31.49	ਨਾਲ	3.66	15.26	ਪੰਜਾਬ	0.99	7.96	ਮੈਂ	0.85	4.54

Table 2: Mean and variance of the top 50 words in descending order of variance.

The probability of the word W_i in each document D_j is denoted as $P_{ij.}$ This probability is calculated as frequency of the word in the document divided by the total number of words in the document. The mean of probability (MP) among the documents for each word is calculated as per the formula:

$$MP(W_i) = \frac{\sum_{1 \leq j \leq n} P_{ij}}{N}$$

The table 3 shows the top 50 words with mean probability in descending order. The stop words should indicate the high value of mean probability.

Word	MP	Word	MP	Word	MP	Word	MP	Word	MP
ਆ	0.0722	ਜਾ	0.0185	ਇਸ	0.0108	ਲਈ	0.0063	ਉਹ	0.0042
ਚ	0.0502	ਕਿ	0.0179	ਹਾਂ	0.0099	ਇਕ	0.0063	ਉਨ੍ਹਾਂ	0.0039
ਦੇ	0.0373	ਹੀ	0.0179	ਅਤੇ	0.0098	ਜਾਂ	0.0056	ਜੋ	0.0037
ਦੀ	0.0340	ਵਿਚ	0.0168	ਹਨ	0.0090	ਉਸ	0.0055	ਸਨ	0.0034
ਦਾ	0.0305	ਕੇ	0.0144	ਨਾਲ	0.0088	ਹਰ	0.0055	ਕੀਤਾ	0.0032
ਤੇ	0.0287	ਹੋ	0.0140	ਤੋਂ	0.0087	ਇਹ	0.0051	ਲੈ	0.0032
ਨਾ	0.0238	ਨੇ	0.0138	ਤਾਂ	0.0072	ਨਹੀਂ	0.0047	ਵਾਰ	0.0032
ਹੈ	0.0231	ਸਿੰਘ	0.0135	ਆਪ	0.0071	ਕਰਨ	0.0047	ਆਪਣੇ	0.0027
ਕਰ	0.0194	ਵੀ	0.0113	ਜੀ	0.0068	ਗਿਆ	0.0045	ਵੱਲੋਂ	0.0026
ਨੂੰ	0.0194	ਸੀ	0.0112	ਪਰ	0.0065	ਦੀਆਂ	0.0044	ਜਾਣ	0.0026

Table 3: Mean probability of stop words in descending order

The stop words are also expected to have stable distribution in the documents. The stability of distribution of words is measured using the variance of probability (VP). The variance is calculated using the formula:

$$VP(W_i) = \frac{\sum_{1 \leq j \leq n} (P_{ij} - \bar{P_{ij}})^2}{N}$$

5. Aggregation of VP and MP

The probability of a word to be a stop word is inversely proportional to its variance of probability. The MP and VP of words together can be used to define a decision variable d_j as:

$$d_i = \frac{MP(w_i)}{VP(w_i)}$$

The higher value of this decision variable, the more the chances for the word to be a stop word. The table 4 shows the MP, VP and d_i values of the words in descending order for top 50 words. These words have the higher chances to be included in the stop word list.

Word	MP	VP	di	Word	MP	VP	di
ਕਰਨ	0.004655	0.00003	170.8099	ਚ	0.050215	0.00043	116.645
ਨੂੰ	0.019397	0.00011	169.6508	ਨੇ	0.013843	0.00012	112.8684
ਤੇ	0.028714	0.00017	164.4231	ਲਾ	0.019218	0.00017	112.058
ਤੋਂ	0.008657	0.00005	163.575	ਜਾ	0.018485	0.00017	111.944
ਵੀ	0.011327	0.00007	161.6505	ਦੀ	0.033965	0.00031	109.367
ਲਈ	0.00631	0.00004	161.0014	ਕਿ	0.017932	0.00016	108.8011
ਨਾਲ	0.008757	0.00006	152.8026	ਆਪ	0.007131	0.00007	107.2144
ਪਰ	0.006504	0.00004	152.4375	ਹੀ	0.01789	0.00017	106.5345
ਇਹ	0.005077	0.00003	149.107	ਜਾਂ	0.005625	0.00005	105.2827
ਹੋ	0.013994	0.00010	145.5602	ਅਤੇ	0.009802	0.00009	103.8446
ਕੇ	0.01443	0.00010	143.1486	ਹਨ	0.009006	0.00009	103.1966
ਕਰ	0.019449	0.00014	141.9663	ਏ	0.009393	0.00009	100.2443
ਦੇ	0.037308	0.00027	138.827	ਦਾ	0.030549	0.00031	97.16349
ਨਹੀਂ	0.004717	0.00003	137.038	ਸੀ	0.011207	0.00012	97.00522
ਇਸ	0.01081	0.00008	136.1846	ਆ	0.072234	0.00076	94.90341
ਇਕ	0.006275	0.00005	135.5775	ਵਿਚ	0.016784	0.00018	93.50988
ਤਾਂ	0.007197	0.00005	135.4527	ਪਾ	0.009012	0.00010	92.53396

ਨਾ	0.023791	0.00018	130.7256	ਟੀ	0.005508	0.00006	86.03452
ਗਿਆ	0.004457	0.00003	130.4762	ਡੀ	0.004801	0.00006	85.88009
ਦੀਆਂ	0.004428	0.00003	128.3804	ਡਾ	0.005147	0.00006	81.1695
ਈ	0.025156	0.00020	124.0861	ਉਸ	0.005538	0.00007	79.63611
ਹਰ	0.005458	0.00004	122.642	ਜੀ	0.006777	0.00009	78.42792
ਕੀ	0.01162	0.00009	122.459	ਹੈ	0.023069	0.00033	69.14594
ਕੁ	0.005635	0.00005	119.9874	ਸ	0.17538	0.00340	51.62368
ਹਾਂ	0.009913	0.00008	119.2854	ਸਿੰਘ	0.013523	0.00064	21.25985

Table 4: Mean Probability (MP), Variance of Probability (VP) and Decision variable ($d_{i)}$ in descending order for top 50 words

6. Results

The results from the frequency count as well as frequency distribution of the words are combined to generate a final stop words list. This compilation is done as per the Borda's Rule[9]. A sorted list from each method is obtained first. The words in each list are ranked as per their position in the given descending ordered list. Finally a sum of the ranks obtained by each word in all the lists is obtained to get its final score. A descending ordered list is prepared as per the final rank of the words. The words towards the top of the list are assumed to have a higher probability of being a stop word.

7. Conclusion

Punjabi stop words list is an important tool required during the text classification process. In this paper, we presented an automated tool for the construction of Punjabi stop words list. The stop word lists prepared by two different approaches were nearly similar, however the aggregation of the two lists changed the order of some words in the list. The revised order will surely benefit the algorithms making linear searches into this list as there are high chances of getting a hit towards the beginning of the list. Since we have used only statistical methods for the construction of the list, this method can further be applied to virtually any language in this world.

References

1. C. J. van Rijsbergen, Information Retrieval, 2nd Edition, Butterworths, London, 1979. URL http://www.dcs.gla.ac.uk/Keith/Preface.html
2. K. Zipf, Selective Studies and the Principle of Relative Frequency in Language, Cambridge, MA; MIT Press, 1932.
3. Lo, R. T.-W., He, B., and Ounis, I. (2005). Automatically building a stop word list for an information retrieval system. In Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR).
4. Sinka, M. P. and Corne, D. (2003a). Evolving better stoplists for document clustering and web intelligence. In HIS, pages 1015–1023.
5. Vishal Gupta and Gurpreet Singh Lehal (2011) Preprocessing Phase of Punjabi Language Text Summarization . Information Systems for Indian Languages Volume 139 of the series “Communications in Computer and Information Science” pp 250-253
6. Feng Zou et al. Automated construction of Chinese Stop word list. Proceedings of the 5th WSEAS International Conference on Applied Computer Science, Hangzhou, China, April 16-18, 2006 (pp1010-1015)
7. R. K. Belew. Finding Out About. Cambridge University Press, 2000.
8. W. Francis. Frequency Analysis of English Usage: Lexicon and Grammar. Houghton Mifflin, 1982.
9. R.B. Myerson, Fundamentals of social choice theory, Discussion Paper No.1162, 1996