A Rule-Based Approach to Identify Stop Words for Gujarati Language

Rajnish M. Rakholia and Jatinderkumar R. Saini

Abstract Stop words removal is an important step in many natural language processing (NLP) tasks. Till now, there is no standardized, exhaustive, and dynamic stop word list created for documents written in Indian Gujarati language which is spoken by nearly 66 million people worldwide. Most of the existing stop words removal approaches are file or dictionary based, wherein a hard-coded static, nonstandardized, and individually created list of stop words is used. The existing approaches are time consuming and complex owing to file or dictionary preparation by collecting possible stop words from a large vocabulary, complex framework and a morphologically variant Gujarati document. Even the other proposed approaches in the literature are also very restricted due to their dependence on word-length, word-frequency, and/or training data set. For the first time in scientific community worldwide, this paper proposes a dynamic approach independent of all factors namely usage of file or dictionary, word-length, word-frequency, and training dataset. An 11 rule-based approach is presented focusing on automatic and dynamic identification of a complete list of Gujarati stop words. Extensive empirical evidence has been presented through deployment of proposed algorithm on nearly 600 Gujarati documents, categorized into routine and domain-specific categories. The respective results with 98.10 and 94.08% average accuracy show that the proposed approach is effective and promising enough for implementation in NLP tasks involving Gujarati written documents.

Keywords Gujarati · Natural Language Processing (NLP) · Rule-based approach · Stop word

R.M. Rakholia (✉)
School of Computer Science, R K University, Rajkot, Gujarat, India
e-mail: rajnish.rakholia@gmail.com

J.R. Saini
Narmada College of Computer Application, Bharuch, Gujarat, India
e-mail: saini_expert@yahoo.com

S.C. Satapathy et al. (eds.), *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications*, Advances in Intelligent Systems and Computing 515, DOI 10.1007/978-981-10-3153-3_79

1 Introduction

Natural language processing is a field of computational linguistic and the goal of NLP is to analyze, understand, and generate human understandable language. But this goal is not easy to reach, because different language has own grammatical structure. To understand dependency among words and sentences, ambiguity of word, and how to link those concepts together in a meaningful say it is challenging task in NLP [1].

1.1 Stop Words

Stop word is a word which has less significant meaning than other tokens. Identification of stop words and its removal process is a basic preprocessing phase in NLP and data mining applications. For any NLP tool there is no single universal list of stop words used for a specific language, because stop words list is generally domain specific [2].

1.2 Diacritics

Diacritic is a mark that is used to change the sound value of the character. Diacritic mark could be identifying by unique UTF-8 value. And by using with any consonant in Gujarati language, it is possible to produce multiple meaning. A list of diacritic marks presented in Table 1 and is further elaborated based on wide and rare usage of the concerned diacritic [3].

Table 1 Diacritics for Gujarati document

Widely Used Diacritics in Gujarati Document				
ા (U+0ABE)	િ (U+0ABF)	ી (U+0AC0)	ુ (U+0AC1)	ૂ (U+0AC2)
ે (U+0AC7)	ૈ (U+0AC8)	ો (U+0ACB)	ૌ (U+0ACC)	ં (U+0A82)
ૃ (U+0AC3)	ૅ (U+0AC5)	ૉ (U+0AC9)	્ (U+0ACD)	ઁ (U+0A81)
Rarely Used Diacritics in Gujarati Document				
ઃ (U+0A83)	ૄ (U+0AC4)	઼ (U+0ABC)	ૢ (U+0AE2)	ૣ (U+0AE3)

1.3 Gujarati Language

Gujarati is an official and regional language of Gujarat state in India. It is 23rd most widely spoken language in the world today, which is spoken by more than 46 million people. Approximately 45.5 million people speak Gujarati language in India and half million speakers are from outside of India that includes Tanzania, Uganda, Pakistan, Kenya and Zambia. Gujarati language is belongs to Indo-Aryan language of Indo-European language family and it is also closely related to Indian Hindi language [4].

1.4 Unicode Transformation Format (UTF)

Unicode Transformation Format (UTF) is a character set [5] which is used to display the character written in Indian languages. We have used 8-bit encoding system to process Gujarati written document which is not possible to display each character using American Standard Code for Information Interchange (ASCII). There are many representations of UTF including utf8, utf16 and utf32 in which UTF-8 is widely used in web technology and mobile application for Indian languages.

2 Related Works and Existing Approaches

Pandey and Siddiqui [6] prepared a list of stop words for Hindi language based on its frequency and some manually operations. For experiment they used EMILLE corpus dataset, precision, and recall was used for evaluation. By removing stop words from raw content, it is possible to improve the accuracy of retrieval [6].

Kaur and Saini [7] they presented natural language processing approach to identify stop words in Panjabi literature in which they concentrates on poetry and other news articles for data collection. They identify 256 stop words from selected category and released for public use [7]. Kaur and Saini [8] described pre-processing phases for Punjabi language, in which, they have manually analyzed the data set (Punjabi text documents) and identified 1,500 stop words. High-frequency terms occurring in document, they have also considered stop word [8]. Kaur and Saini [9] they have provided enhanced understanding of stop words in Panjabi language based on Part-of-speech tagging. They constructed data set from different five categories of Panjabi literature: natures, romantic, religious, patriotic and philosophical, are manually populated with 250 poems. They prepared 256 stop words manually, due to unavailability of Punjabi stop words in public domain [9].

Thangarasu and Manavalan [10] developed stemmer for Tamil language; stemming algorithm pay important role to create stop words list. They created a list of tokens which is available in text corpus. After shorting that list and based on token frequency they prepared stop words list and other words to be discarded [10].

Yao and Zen-wen [11] created list of 1289 Chinese-English stop words by combining domain-specific stop words with list of classical stop words [11]. For Mongolian language, [12] used entropy calculation to create stop words list. They calculate entry for each word that is available in initial created stop words list. To prepare final stop words list, they combine this result with Mongolian part-of-speech [12]. Alajmi et al. [13] have used statistical approach to generate stop words list for Arabic language [13].

Chauhan et al. [14] presented stemmer for Gujarati language by using rule-based approach to improve Information Retrieval System. They used Gujarati news paper corpus for experiment purpose and created list of 280 stop words based on a word which is frequently occurring and it is less importance in document [14]. Joshi et al. [15], presented stop word elimination approach for information retrieval (IR) of Indian Gujarati language to improve mean average precision (MAP). They have collected data from FIRE corpus, based on their experiment, they constructed list of 400 words which is less importance and extensively used in Gujarati language. They created 282 stop words list from constructed list by analyzing and manually inspection by linguistic expert [15]. Rakholia and Saini [16] they study and analyzed different stemmer algorithms and pre-processing approaches are available for Gujarati language to process Gujarati written document. Through of their literature they found that, stop words removal is important pre-processing step in natural language processing application [16].

Based on this detailed literature review of the most relevant research works found in research community, our analysis based on stop words identifying process for Gujarati written document, it has been found by us that most of the researchers have obtained average accuracy for training and testing phase for Gujarati stop word identification at 85 and 67%, respectively. This motivated us for the presented research work as there is no effective stop word identification method or approach developed for Gujarati written document, which can yield a performance enough to make it practically acceptable in real world.

3 Our Approach

We have used rule-based approach to identify stop words from Gujarati written document. We have not considered the length of the word to identify the stop word because Gujarati document can be written using consonants, vowels and diacritics signs as well. It is noteworthy to mention here that the length of the stop word found by methods used by other researchers hence is dependent on and influenced by the usage of diacritics as well. To design and implement a length independent approach, we have deployed the usage of the fact that each diacritic mark in written

Gujarati document considers a single character. Also, from linguistic computational perspective, each diacritic mark has a unique UTF-8 hexadecimal value.

Following rules are applied to identify stop words appearing in Gujarati document

Rule 1: All single consonant or vowel words, with or without diacritics, were considered stop word and eliminated, except only **{મા, ચા, બા, દી, પી and ગૌ}.**

For instance: With diacritics: **{તે, જે, જી, તો, છે, મે, એ, કે etc. }**
Without diacritics: **{ન, પ, સ, ક, વ, દ, છ, ઈ, ઉ etc.}**

Rule 2: A word that contains three regular Gujarati characters other then diacritic sign, if a word is terminated with **"થી"** and if a middle character has "ા" diacritic sign and first character has either "ા" or "ે" sign, then it was considered stop word and eliminated.

For instance: **{તારાથી, કેનાથી, જેનાથી, તેનાથી, આનાથી, કોનાથી, એનાથી, હોવાથી etc.}**

If a word that contains two regular Gujarati characters other then diacritic sign and if word is terminated with **"થી"**, then it was considered stop word and eliminated.
For instance: **{આથી, તેથી, જેથી, નથી etc.}**

Rule 3: A word that contains only two regular Gujarati characters other then diacritics sign and if word is terminated with **"વુ" or "વો",** then it was considered stop word and eliminated.

For instance: **{આવુ, કેવુ, જેવુ, લેવો, કેવો, તેવુ, etc.}**

Rule 4: A word that contains only two regular Gujarati characters other then diacritics sign and the word is terminated with **"ની"** and word does not start by using this three diacritics sign {ા, િ, ી} then it was considered stop word and eliminated. Because in most cases, these three diacritic signs {ા, િ, ી} are used to make proper nouns (e.g., name of girls) in Gujarati language.

For instance: **{તેની, એની, જેની, દેની, લેની, કોની etc.}**

Rule 5: A word that contains only two regular Gujarati characters other then diacritics sign and if word is terminated by **"હા"** with at least one diacritic sign and does not start with **"થા"**, then it was considered stop word and eliminated.

For instance: **{કોણે, એણે, તેણે, જેણે, ગણે, etc.}**

Rule 6: A word that contains only two regular Gujarati characters other then diacritics sign and if word is terminated by **"નુ"** and starting character has only "ે" diacritic sign, then it was considered stop word and eliminated.

For instance: **{કેનુ, જેનુ, તેનુ, દેનુ, એનુ, સેનુ, etc.}**

Rule 7: A word that contains only two regular Gujarati characters other then diacritics sign and if word is terminated by **"ઈ"** and first character either does not contain diacritic sign or have only "ે" diacritic sign, then it was considered stop word and eliminated.

For instance: {કઇ, ગઇ, લઇ, જઇ, કેાઇ, **etc.**}

Rule 8: A word that contains only two regular Gujarati characters {હ, ત } other then diacritic sign and last character has at least one diacritic sign when first character has "ી" or "િ" sign, then the word under consideration was treated as a stop word and eliminated. Using this rule, it was also possible to identify past tense sentences written in Gujarati language.

For instance: {હતુ, હતી, હેાત, હતો **etc.**}

Rule 9: A word that contains only two regular Gujarati characters other then diacritics sign and if word is terminated with **"ને",** then it was considered stop word and eliminated.

For instance: {કોને, તને, એને, કેને, અને, જેને, શાને, એાને, રેને, તેને **etc.**}

Rule 10: A word that contains two regular Gujarati characters other then diacritic sign and if word is terminated with **"મ"** and first character contained at least one diacritic sign except "ે", then it was considered stop word and eliminated.

For instance: {જેમ, કેમ, તેમ, એમ, **etc.**}

Rule 11: A word that contains two or three regular Gujarati characters other then diacritic sign and if word is terminated with **"તેા" or "તે"**, then it was considered stop word and eliminated.

For instance: {થતે, એતો, જતો, કેતો, વખતે, આવતે, જતે **etc.**}

4 Comparison with Other Approaches

Almost researchers have created stop words list for Indian Gujarati language by manually inspection of linguistic expert and based on words frequency. A list of existing approaches that are used for Indian language to identify stop words is presented in Table 2.

Other than these approaches, statistical approach is also used to generate stop words list. In almost all existing approaches, first step is frequency calculation for each word. But in many cases a word that has high frequency with significant meaning in document, but it cannot be consider as stop word. Second, many researchers have used statistical approach for English language and they achieved good accuracy, because many stop words in English language does not have multiple form, for instance: "any", "is," "a," "the," "an." But for the Indian Gujarati

Table 2 Existing approaches

S. No.	References	Dataset/corpus	Existing approach	Language and no. of stop words
1	Pandey and Siddiqui [6]	EMILLE	Frequency-based and manual operations	Hindi (Not Provided)
2	Kaur and Saini [7]	Dataset from Poetry and news article	Manual inspection	Panjabi (256)
3	Kaur and Sharma [8]	Panjabi text document	Frequency based	Panjabi (1500)
4	Thangarasu and Manavalan [10]	Text corpus (Not Provided which corpus used)	Frequency based	Tamil (Not Provided)
5	Chauhan, Patel and Joshi [14]	Gujarati news paper corpus	Frequency based	Gujarati (280)
6	Joshi H et al. [15]	FIRE Corpus	Manual inspection	Gujarati (282)
7	Proposed Approach	Routine Gujarati and domain specific	Rule-based	Gujarati (Dynamic)

language statistical approach will lead to the loss of accuracy because single stop words has multiple form, for instance: "હતુ", "હતી", "હોત", "હતા".

4.1 *Precise Benefits of Proposed Approach Over Existing Approaches*

The research works found in the related literature are based on training dataset and/or the length of the word. The proposed approached is free from the length of the word as well as the requirement of the training data set. It is noteworthy to mention that deploying a training dataset often leads to biased training of the system, more so in absence of availability of a standard text corpus for a resource scarce language like Gujarati. The proposed approached is hence free from machine learning based techniques. The proposed approach is also, hence, free from the risk of getting obsolete with time.

4.2 *Known Limitations of Proposed Approach*

The proposed work, in its present state, "will not perform well" only in case of stop words that contain more than three characters. It will also "not perform well" with specific words that belong to a peculiar domain. Still, two points are worth

mentioning here. Firstly, the phrase "will not perform well" here should be taken with a pinch of salt as the only detrimental thing from the system will be a slight reduction in the accuracy. Second, the probability of peculiar domain stop word identification is very less, more so during the usual text processing and natural language processing tasks for any language, again much more so for resource scarce language Gujarati. In neither case, the proposed rules prove to be injurious enough preventing the system from wide implementation and its acceptability with good reputation in the scientific community.

5 Empirical Setup and Results

Indeed there is no a priori definition of stop words and their handling is governed by the domain and application area they are used for. Still, the NLP tasks like machine translation (MT), POS-tagging, and classification make use of general stop-word removal phase. To say "general stop-word" removal emphasizes on the fact that there are words with high frequency and their removal helps in faster processing as well as also helps in dimension reduction in terms of space requirement. This paper does neither intend to highlight the domain or application area in which stop words should be removed, nor does it focus on the number of stop words to be removed. The scientific literature of natural language processing has many instances of stop-word removal. This is true for Gujarati language, other Indo-Aryan languages as well as various International languages. This paper emphasizes on the fact that if the stop words have to be removed for Gujarati documents, there is no need to implement word frequency-based approach, word-length based approach, or manual inspection. Exploiting the morphological structure and symmetry of Gujarati stop words, this paper proposes a rule-based approach for stop word removal from Gujarati documents. This approach could be used anywhere where general (i.e., non-application and non-domain specific) removal of stop words is required. Even for cases where application and domain-specific removal of stop words is required for extrinsic evaluation of any system, the proposed "generic" rules could be applied before implementing domain and application specificities. As the proposed rules could be applied anywhere where removal of stop words is required, we term them 'generic'.

This section described the source of data collection for empirical implementation of the proposed rules. The system was implemented using Java Server Pages (JSP) technology and the results follow.

5.1 Data Sets

The data was collected randomly from multiple free Gujarati websites, to avoid the bias of a single website on the proposed work. For experimental purpose,

373 documents were prepared for routine Gujarati document and each document contained more than 400 words. We also prepared 224 documents for domain-specific (medical and engineering) categories and each document in these categories contained more than 275 words.

5.2 Results

In Gujarati language, there is no automated tool readily available to calculate the accuracy. Hence, we had to manually go through each document and evaluate the performance of the system. The obtained results on accuracy were recorded side by side. The average accuracy of routine Gujarati written documents was obtained at 98.10%. Similarly, for domain-specific medical and engineering categories, the obtained average accuracy was 94.08%. We also pondered on the reasons of getting less accuracy for routine Gujarati written documents and found that the reason is the presence of stop words containing more than three characters. Similarly, the non-availability of 100% accuracy in case of domain-specific categories owes to the presence of peculiar domain biased words. The average accuracy of routine Gujarati written documents is greater than the average accuracy of specific domain category documents by 4.02% because of presence of many domain-specific words, in such documents, which were not identify by any rules.

6 Conclusion and Future Work

We have presented an effective approach to accurately identify and eliminate a high percentage of the stop words in the Gujarati written documents. The proposed work used rule-based approach to identify stop words dynamically. The average accuracy for routine Gujarati written documents was obtained at 98.10% and for the specific domain (Medical and Engineering), we got 94.08% accuracy. We advocate that these results are reproducible on other large corpuses of routine Gujarati written documents as well. We propose and strongly claim that this approach is more efficient than any other existing approaches, which are available for identification of stop words from Gujarati written documents. The approach to finding stop words that presented here is currently limited in its applicability only for the word that contains more than three characters and for the word that belongs to a specific domain. This is our focus for future work. The proposed approach can be well applied as a preprocessing step for many NLP tasks including text classification, information retrieval, as well as document clustering, to name a few.

References

1. Microsoft Research, Natural Language Processing [online] available: http://research.microsoft.com/en-us/groups/nlp/ [Feb 10 2016].
2. Wikipedia, Stop Words Basic [online] available: https://en.wikipedia.org/wiki/Stop_words [Feb 5, 2016].
3. Rakholia R and Saini J, "The Design and Implementation of Diacritic Extraction Technique for Gujarati Written Script Using Unicode Transformation Format", Proceeding of ICECCT, IEEE, 2015, pp. 654–659.
4. UCLC, Gujarati Language [online]: http://www.lmp.ucla.edu/Profile.aspx?LangID=85&menu=004 [Feb 10 2016].
5. The Unicode Consortium, USA, The Unicode Standard [Online]. Available: http://www.unicode.org/standard/standard.html [December 15, 2015].
6. Pandey A and Siddiqui T, "Evaluating Effect of Stemming and Stop-word Removal on Hindi Text Retrieval", Proceedings of the First International Conference on Intelligent Human Computer Interaction, Springer, 2009, pp. 316–326.
7. Kaur J and Saini J, "POS Word Class based Categorization of Gurmukhi Language Stemmed Stop Words", accepted for publication in the proceedings of International Conference on ICT for Intelligent Systems (ICTIS-2015), supported by ACM, CSI and Information Security Research Association and held during November 28–29, 2015, Ahmedabad.
8. Kaur R and Sharma S, "Pre-processing of Domain Ontology Graph Generation System in Punjabi", International Journal of Engineering Trends and Technology, Volume 17 Number 3 – Nov 2014, pp. 141–146.
9. Kaur J and Saini J, "A Natural Language Processing Approach for Identification of Stop Words in Punjabi Language", published in International Journal of Data Mining and Emerging Technologies; ISSN: 2249-3212 (eISSN: 2249-3220); Indian Journals, New Delhi, India; vol. 5, issue 2, November 2015; pages 114–120.
10. Thangarasu M and Manavalan R, "Design and Development of Stemmer for Tamil Language: Cluster Analysis", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, pp. 812–818, July 2013.
11. Yao Z and Ze-wen C, "Research on the construction and filter method of stop-word list in text Preprocessing", Fourth International Conference on Intelligent Computation Technology and Automation, 2011.
12. Zheng G and Gaowa G, "The Selection of Mongolian Stop Words", IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS), 2010.
13. Alajmi A. et al., "Toward an ARABIC Stop-Words List Generation", International Journal of Computer Applications, Volume 46– No. 8, May 2012.
14. Chauhan K, Patel R and Joshi H "Towards Improvement in Gujarati Text Information Retrieval by using Effective Gujarati Stemmer" Journal of Information, Knowledge and Research in Computer Engineering, Nov 12 TO Oct 13, Volume – 02, Issue – 02, Page 218.
15. Joshi H. et al, "To stop or not to stop — Experiments on stopword elimination for information retrieval of Gujarati text documents" Engineering (NUiCONE), 2012 Nirma University International Conference on, 6–8 Dec. 2012, Page 1–4, IEEE.
16. Rakholia R and Saini J, "A Study and Comparative Analysis of Different Stemmer and Character Recognition Algorithms for Indian Gujarati Script", published in International Journal of Computer Application (IJCA); Digital Library ISSN: 0975-8887; ISBN: 973-93-80883-64-4; Foundation of Computer Science, USA; vol. 106, issue 2; November 2014; pages 45–50; DOI: 10.5120/18496-9558