Automatic Generation of Stopwords in the Amharic Text

Sileshi Girmaw Miretie
CS-IT Dept
Symbiosis Institute of Technology,
Symbiosis International University,
Pune, India

Vijayshri Khedkar
CS-IT Dept
Symbiosis Institute of Technology,
Symbiosis International University,
Pune, India

ABSTRACT
For the retrieval of information from documents of different natural languages, pre-processing of the document is the main task. During pre-processing, words which occur too frequently and have little semantic in the document should be identified. Such words are called Stopwords. Stopwords list for different world languages like English, Chinese, Hindi, Arabic Sanskrit etc. are identified. But as I long as I know there is no standard method to identify these words for the Amharic language. In this paper, we proposed the automatic identification of Stopwords for the Amharic text by an aggregate based methodology of words frequency, inverse document frequency, and entropy value measure. Available works on Stopwords identification techniques are based on static or dictionary based Stopwords lists. This method inefficient and very expensive and it is a time-consuming task as the searching process takes a long time. The proposed work will overcome these problems using aggregated methods of both frequency measures and entropy measures of words in the Amharic text for the automatic Stopwords identification.

Keywords
Natural language processing, information retrieval, document pre-processing, Stopwords, Amharic

1. INTRODUCTION
pre-processing is the vital task of natural language processing, information retrieval, Artificial intelligence, web mining, text classification, stemming etc. for a given document. Identification and removal of stop words from given document is the main step during the preprocessing task. Stopwords are general words which have little meaning if they are used separately, are frequently occurring words in the document and they are useful for the structure of the language not for the semantics of the language.

Different techniques are used to identify these words from the document by different researchers. Dictionary-based approach, supervised approach for distinguishing between keywords and stopwords using probability distribution, automated algorithm based on the frequency of words, deterministic finite automata entropy-based approach which is a major metric for the level of information a word contains in documents, a revised statically approach which is based on frequency count and distribution of words in different documents and similarity function and part of speech information are some of the techniques to identify general and domain-specific stopwords from documents. However, except static or Dictionary based approach, automatic generation of stopwords for the Amharic text are not available.

In this paper, we are going to identify stopwords automatically from in Amharic text using the aggregated-based technique. One is bases of word frequency and the other is based on entropy measures of words in the given documents of the Amharic text. Amharic is Ethiopian national language which has about thirty-two alphabets. More than ninety million people speak this language as native and second language. The sources data for this research are Amharic newspapers, magazines and known blogs which are considered as they are written in correct language structure. And this technique enables us to identify the stop-word lists without affecting the content of information the original document before removing the non-informative words. Identification of these stopwords enables the language users to retrieve information fast and makes the language more powerful for information processing. For example, Amharic words like "ይጠበቃል" (expected) "አስታውቀዋል" (declared), ብለዋል (said), "ግን "(but), "ነው" (is), "ነበር" (was) etc. are considered as stopwords.

The remainder of these paper has five parts. Section II is an overview of related works. Section III is a brief explanation of our method for stopwords identification. Section IV is an experiment of the method. finally, the last section is about the conclusion and future work.

2. RELATED WORK
In natural language processing and related fields, various researchers have been done on the idea of identification and removal of stopwords different languages. Automated Stopwords identification is the most efficient and widely used method with a little or no intervention of manual methods. Jaideep Singh et al [1] used automatic stopwords identification algorithm for the Sanskrit language and some manual intervention is used by the language expert, and then call the method hybrid. They calculated the frequency of words from the input text and they also used some words from the dictionary to identify the stop list. Asubiaro, Toluwase V, [3] used an entropy-based algorithm to identify stopwords for Nigerian Yoruba language text. A word whose entropy is greater than 0.6 but not a noun was considered as a Stopword. Walaa Medhat et al [4] generated stopwords list for the Egyptian dialect for online social network data to investigate the effects or removal of stopwords from the text for the sentiment analysis (SA) task using frequency the frequency of words from the input Egyptian dialect. Mohammed-Ali Yaghoub-Zadeh-Fard et al [5] generated stopwords list for Persian language Information retrieval system based on similarity function and POS information using the aggregated method of part of speech and statistical features of stopwords. Vijayarani S, et al [7] used Zipf's Law (Z method) for creation of stop-words. Rakholia and Saini [8] have presented a rule based approach to dynamically identify stop words for Gujarati language. Vandana Jha et al [9] developed an algorithm to remove stopwords from the Hindi text based on Deterministic finite automata. The algorithm also tested on 200 documents and succeeded 99% accuracy and time

efficiency. Saini and Rakholia [13] have presented an analytic in-depth report on continent and script-wise divisions-based statistical measures for stopwords lists of various international Languages. A. Alajmi et al [19] generated stop-words for the Arabic language using a statistical approach.1002 documents with over 700,000 words were tested and they achieved about 90% general accuracy. El-Khair,et al [25] conducted research on the effectiveness of three stop words lists for Arabic Information Retrieval---General Stoplist, Corpus-Based Stoplist, Combined Stoplist ---were investigated in this study. Three popular weighting schemes were examined: the inverse document frequency weight, probabilistic weighting, and statistical language modelling. The Idea is to combine the statistical approaches with linguistic approaches to reach an optimal performance, and compare their effect on retrieval.

3. PROPOSED WORK

Different approaches are used by researchers to generate and remove stopwords from the documents of different languages of the world. Some of these methods are Dictionary based approach, supervised approach using probability distribution, automated algorithm based on the frequency of words, deterministic finite automata entropy measures approach for the contents of information of a word in the document, a revised statically approach which is based on term frequency and distribution of words in different documents and studying part of speech are some of the techniques to identify general and domain-specific stopwords from documents. Amharic is a national/working language having its own grammar and syntax structure. However, as long as I know, there is no general list of stopwords for the Amharic language. Stopwords in Amharic should have the following properties.

•They are non-informative words if they are used alone.

•They occur frequently in documents.

•Important for the structure of the language not important for the semantics purpose.

•Most of the time they can be adjectives, pronouns, Articles.

•General words for the language and are not domain specific.

In this paper, we are going to identify stopwords automatically from the Amharic text using the aggregated-based technique. One is bases of word frequency; inverse document frequency and the other is based on entropy measures of words in the given documents. The data inputs for this research are from magazines, newspapers, and blogs written with the proper structure of the language.

3.1 Term frequency

The count or number of times each term (t) occurs in each document (d) is called its term frequency. From the lists of words that we get from magazines, newspapers, and blogs as inputs, we can calculate the frequency of each word in the documents and it shows some measure of term density in a document. This measure is very important to determine the most relevant document to the query terms from a set of text documents. The best way to apply is by eliminating the documents that do not contain all the terms we need. So to further distinguish, we have to count the number of times each word is coming in a document and then sum up them together. This sum is what we call "term frequency". Thus, terms with high frequency are considered as less informative terms in the document. And most researchers used this measurement for the stopwords list identification for different world languages. Term frequency a term can be defined as

$$\boldsymbol{tf} = (\boldsymbol{tf}, \boldsymbol{d}) / (\sum \boldsymbol{ft}, \boldsymbol{d},)$$

Where,

$\boldsymbol{tf}, \boldsymbol{d}$ is Term frequency in a document and $\sum \boldsymbol{ft}, \boldsymbol{d}$ total word number of terms of documents

3.2 Inverse Document Frequency(idf)

Inverse Document Frequency is the measure of the uniqueness of a term. It shows whether a term is common or rare in the document. In the computation of term frequency, we have considered all the terms are important. In the Amharic text, although you all know that few terms like "እና", "ነው", and "ግን" appear a lot of times in the document but they are having little importance. Hence, we must lower the weight of frequent occurring terms and increase their rareness. The inverse document frequency for any given term is defined as,

$$\text{idf} = \log\left(\frac{\text{N documents}}{\text{N documents containing term}}\right)$$

For example, Here Let's Consider that total number of documents is 10,000.

Table 1. example of inverse document frequency

Amharic Terms	Total number of Documents containing the term
ነው	10000
ነበር	5000
መምህር	100

From the above table, we can understand that the term "ነው" appears in every document. But it shows that it provides no value in the document as it is present in every document. If we look at the term "መምህር", it is present in 100 documents. Clearly, from this information, we can understand that this term is having high importance in the documents in which it is present.

3.3 Entropy measures

In information theory, word information bearing capacity correlates the randomness of a word. Shanon [22] suggests that a randomness measure of a word is called entropy. Then, words with high randomness and are also low entropy words are considered as very informative. Since stopwords are less. Informative they are high entropy words. Entropy measures the frequency variance of a given word for multiple documents, i.e. words with very high frequencies in some documents but the low frequency in others will have high entropy. Entropy H (w) of a given word w with respect to a given set of n documents is as follows:

$$H\ (Wj)\ = \Sigma\ Pi, j\ .\ log\left(\frac{1}{Pi, j!}\right)$$

Where,

$$\boldsymbol{Pi}(\boldsymbol{W}) = \frac{\boldsymbol{fi}(\boldsymbol{W})}{\sum_{j=1}^{n} \boldsymbol{fj}(\boldsymbol{w})}$$

$\boldsymbol{fi}\ (\boldsymbol{w})$ = Frequency of word $\boldsymbol{w}$ in document *i*, *n* = number of documents.

The entropy of each word in the dataset will be considered, and the value will be ordered by increasing of entropy to expose the words that have a better probability of being noise words. Finally, by aggregating term frequency, idf, it-idf and entropy measures we can generate most important lists of Amharic stopwords. The following block diagram shows the general structure of the proposed work.

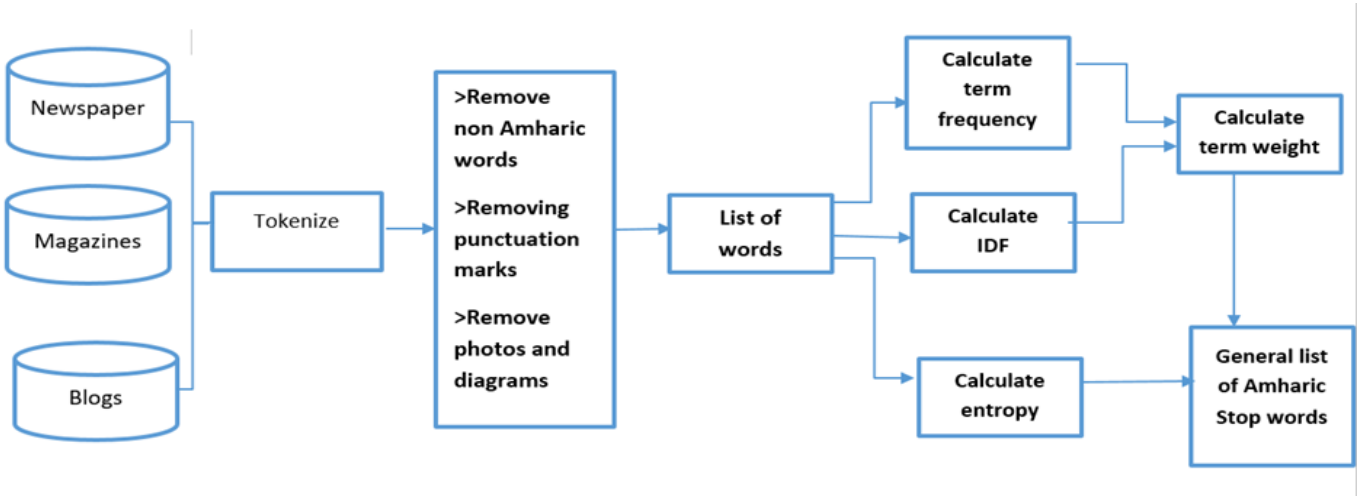

Fig 1: structures of the proposed work

4. CONCLUSION

Stop words list generated for many natural languages of the world. Amharic is also the largest and most important language of Ethiopia .as it's the national language of the country stop words list generation for the language is an important task required for the text processing purposes. In this paper, we proposed to generate Amharic stop words list from the Amharic text. The methodology we are an aggregation high term frequency measure, low term weight measure and high entropy measures. This enables educators, researchers, and language experts etc. to do more on the idea to enhance the language power in various aspects

5. REFERENCES

[1] Raulji, J. K., & Saini, J. R. (2017, January). Generating Stopword List for Sanskrit Language. In Advance Computing Conference (IACC), 2017 IEEE 7th International (pp. 799-802). IEEE.

[2] Raulji, J. K., & Saini, J. R. Stop-Word Removal Algorithm and its Implementation for Sanskrit Language.

[3] Asubiaro, T. V. (2013). Entropy-Based Generic Stopwords List for Yoruba Texts. Entropy, 2(05).

[4] Medhat, W., Yousef, A. H., & Korashy, H. (2015). Egyptian Dialect Stopword List Generation from Social Network Data. arXiv preprint arXiv:1508.02060.

[5] Mohammed-Ali, Y-Z-F., Behrouz, M-B., Saeed, R.,& Saeed, S.(2015, November) PSWG: An automatic Stopword list generator for Persian information Retrieval systems based on similarity function &POS information.2015 international conference on Knowledge Based-engineering and Innovation(KBEI).IEEE.

[6] Saif, H., Fernandez, M., & Alani, H. (2014, October). Automatic stopword generation using contextual

[7] Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. International Journal of Computer Science & Communication Networks, 5(1), 7-16.

[8] Rakholia, R. M., & Saini, J. R. (2017). A Rule-Based Approach to Identify Stop Words for Gujarati Language. Suresh Chandra Satapathy Vikrant Bhateja Siba K. Udgata, 797.

[9] Rakholia, R. M., & Saini, J. R. (2017). Information Retrieval for Gujarati Language Using Cosine Similarity Based Vector Space Model. In Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications (pp. 1-9). Springer, Singapore.

[10] Semantics for sentiment analysis of Twitter. In Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272(pp. 281-284). CEUR-WS. Org.

[11] Puri, R., Bedi, R. P. S., & Goyal, V. (2013). Automated Stopwords Identification in Punjabi Documents. vol, 8, 119-125

[12] Jha, V., Manjunath, N., Shenoy, P. D., & Venugopal, K. R. (2016, January). Hsra: Hindi stopword removal algorithm. In Microelectronics, Computing and Communications (MicroCom), 2016 International Conference on (pp. 1-5). IEEE.

[13] Saini, J. R., & Rakholia, R. M. (2016). On Continent and Script-Wise Divisions-Based Statistical Measures for Stop-words Lists of International Languages. Procedia Computer Science, 89, 313-319.

[14] Sharan, A., & Siddiqi, S. (2014, September). A supervised approach to distinguish between keywords and stopwords using probability distribution functions. In Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on (pp. 1074-1080). IEEE.

[15] Puri, R., Bedi, R. P. S., & Goyal, V. (2013). Automated Stopwords Identification in Punjabi Documents. vol, 8, 119-125.

[16] Na, D., & Xu, C. (2015). Automatically generation and evaluation of Stop words list for Chinese Patents. TELKOMNIKA (Telecommunication Computing Electronics and Control), 13(4), 1414-1421.

[17] Hidayatullah, A. F., & Ma'arif, M. R. (2017, January). Pre-processing Tasks in Indonesian Twitter Messages. In Journal of Physics: Conference Series (Vol. 801, No. 1,

p. 012072). IOP Publishing.

[18] Ferilli, S., Esposito, F., & Grieco, D. (2014). Automatic learning of linguistic resources for stopword removal and stemming from the text. Procedia Computer Science, 38, 116-123

[19] Wilbur, W. J., & Sirotkin, K. (1992). The automatic identification of stop words. Journal of information science, 18(1), 45-55.

[20] Zou, F., Wang, F. L., Deng, X., & Han, S. (2006). Automatic identification of Chinese stop words. Research on Computing Science, 18, 151-162.

[21] Munková, D., Munk, M., & Vozár, M. (2014). Influence of stop-words removal on sequence patterns identification within comparable corpora. In ICT Innovations 2013 (pp. 67-76). Springer, Heidelberg.

[22] Saif, H., Fernández, M., He, Y., & Alani, H. (2014). On stopwords, filtering and data sparsity for sentiment analysis of Twitter.

[23] Alajmi, A., Saad, E. M., & Darwish, R. R. (2012). Toward an ARABIC stop-words list generation. International Journal of Computer Applications, 46(8), 8-13.

[24] Kumar, M., & Vig, R. (2013). Focused crawling based upon Tf-IDF semantics and hub score learning. Journal of Emerging technologies in web intelligence, 5(1), 70-77.

[25] Ospanova, R. (2013). Calculating Information Entropy of Language Texts. World Applied Sciences Journal, 22(1), 41-45.

[26] Shannon, C. E. (1948). A mathematical theory of communication, Part I, Part II. Bell Syst. Tech. J., 27, 623-656.

[27] Harman, D. W. (1986, September). An experimental study of factors important in document ranking. In Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 186-193). ACM.

[28] R. Tsz-Wai, B. He, and I. ―Automatically Building a Stopword List for an Information Retrieval System. ‖ 5th Dutch-Belgium Information Retrieval Workshop (DIR)'05Utrecht, the Netherlands 2005.

[29] Abu El-Khair, I. (2017). Effects of Stop Words Elimination for Arabic Information Retrieval: A Comparative Study. arXiv preprint arXiv:1702.01925

[30] Blanchard, A. (2007). Understanding and customizing stopword lists for enhanced patent mapping. World Patent Information, 29(4), 308-316.