Effective Domain Specific Stopwords Generation for Sinhala Text

S.V.S. Gunasekara[1], Prasanna S. Haddela[2]

Sri Lankans mainly use three languages: Sinhala, Tamil and English in day-to-day life. However, Sinhala is the mother tongue of about 15 million Sinhalese, while it is spoken by 19 million people in total. As a result, they use documents written in Sinhala, to smoothly carry out official government communications with the public. Hence, there is a need for non-English based document analyzing and categorizing systems. To address these issues, researchers turned to a new research area called Text Classification. The removal of stopword list is often viewed as an essential factor of the text classification and it results in providing benefits such as: improve retrieval effectiveness, reduce the size of index and reduce the searching time. Stopword is a term used frequently and offers very little context on their own in any language. Common stopwords in Sinhala are, for example: prepositions (‘**ඉහත**’, ‘**පසුව**’, ‘**පිළිබඳව**’) and conjunctions (‘**හා**’, ‘**ද**’, ‘**විසින්**’, ‘**සඳහා**’). So far stopword list have been developed for almost 50 languages like English, Chinese, Hindi etc. However, there is no widely accepted stopword list for the Sinhala language. Even though, it is easy to use general stopword lists which are implemented already, it is insufficient to use such stopwords in certain applications. For example, in the domain of sports texts, terms like “**ලකුණු**” (score), “**ක්‍රීඩා**” (sports) and “**කණ්ඩායම**” (team) occur almost in every document, and these terms are identified as stopwords based on its importance. In contrary, when criminal data set is considered, the aforementioned stopwords are identified as important words. Generally, the standard stopword list does not cover such domain specific terms. Hence, this paper demonstrates how to generate a domain-specific stopword list from a given data set of Sinhala newspapers. It contained 1000 documents that vary in length and fall into five categories (politics, crime, business, religion and sports). Accordingly, experiments were conducted with seven stopword identification methods (Term Frequency, Normalized Term Frequency, Double Normalized Term Frequency, Document Frequency, Inverse Document Frequency, Normalized IDF, TF*IDF) and classifiers (Maximum Entropy, Naïve Bayes), previously applied to other languages. By using the above methods, a new algorithm for building a domain-specific stopword list is proposed. From this method, the stopwords are observed by threshold value and classified by average F-measure and average accuracy in each category. Depending on the comparative study among seven stopword identification methods, the most effective stopwords identification method can be identified. Similar to previous researches, the Normalized IDF method achieved the best improvementin the accuracy after omitting its unique stopwords including 0.994% value of average F-measure from the given tested dataset. Also, the Maximum Entropy classifier is more sensitive to stopword removal than the Naïve Bayes by comparing F-measure and accuracy of each category. According to the results of study, the effective stopword identification method and classifier can be changedaccording to thesize of the corpus. In future work, to further test this approach, it can be investigated whether other methods and classifiers can be utilized in order to achieve more effective stopword lists for Information Retrieval (IR).

Keywords: Stopwords, Sinhala, Stopword Identification Methods, Threshold Value

[1] Faculty of Graduate Studies and Research, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka

[2] Faculty of Computing, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka