Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 89 (2016) 313 – 319

Procedia
Computer Science

Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016)

On Continent and Script-Wise Divisions-Based Statistical Measures for Stop-Words Lists of International Languages

Jatinderkumar R. Saini[a,c,*] and Rajnish M. Rakholia[b,c]

[a]*Narmada College of Computer Application, Bharuch 392 011, Gujarat, India*
[b]*S. S. Agrawal Institute of Computer Science, Navsari 396 445, Gujarat, India*
[c]*R K University, Rajkot, Gujarat, India*

Abstract

The data for the current research work was collected for 42 different International languages encompassing 3 continents viz. Asia, Europe and South America. The data comprised of unigram model representation of lexicons in the stop-words lists. 13 scripting systems comprising Arabic, Armenian, Bengali, Chinese, Cyrillic, Devanagari, Greek, Gurmukhi, Hanja & Hangul, Kana, Kanji, Marathi, Roman (Latin) and Thai were considered. Based on a comprehensive analysis of statistical measures for Stop-words lists, it has been concluded that Asian languages are mostly self-scripted and that the average number of stop-words in Asian languages is more than those in European languages. In addition to various important and other first research results, a very important inference from the current research work is that the average number of stop-words for any given language could be predicted to be 200.

Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016

Keywords: Function-Words; Language; Natural Language Processing (NLP); Script; Stop-Words.

1. Introduction

Owing to the increased availability of computing power and awareness as well as the need of processing naturally spoken languages by people, the domain of Natural Language Processing (NLP) has come up with a wide scope of research. Often, the research areas of NLP overlap with the other areas of computing including Text Mining (TM) and Artificial Intelligence (AI). These areas of NLP, TM and AI, often even with other areas like Computational Linguistics (CL) and Data Mining (DM), collectively encompass various tasks and sub-tasks including Concept Mining, Information Extraction, Information Retrieval, Stemming, Lemmatization, Search Engine Indexing, Text Summarization, Part-of-speech (POS) Tagging, Wordnet development, Text Analysis and Text Classification, to name a few.

Almost all of the tasks and sub-tasks of NLP and its intra and inter-related areas require that during pre-processing phase or as required, stop-words be removed from the text corpus before further processing. In computing, according to Wikipedia[21], they are the words which usually refer to the most common words in a language. Neither is there a universal list of stop-words, common for all languages, nor is there any standardized list of stop-words for each

[*]Corresponding author. Tel.: +91-9687689708.
E-mail address: saini_expert@yahoo.com

Peer-review under responsibility of organizing committee of the Organizing Committee of IMCIP-2016
doi:10.1016/j.procs.2016.06.076

language. Neither do all NLP tools use such lists nor do all tools prefer not to use them. There have been even instances when even more than one type of list of stop-words has been used by researchers[15,17].

It is notable that Function-words are not synonymous with Stop-words though Function-words may include articles ('the', 'a', etc.), pronouns ('he', 'him', 'she', 'her', etc.), particles ('however', 'then', 'if', 'thus', etc.), etc. Hence, Function-words act as a subset of Stop-words. In the context of current research work it is also important to note that the paper uses the words 'language' and 'script' to mean 'the human spoken natural language' and its 'writing system consisting of alphabets', respectively. The alphabet set for a script may in turn consist of all or some or combinations of consonants, vowels, etc.

The author believes that the language does not become rich if it has a long list of stop-words. Rather, such a language will turn out to be a non-good choice from the perspective of NLP algorithms. It is so because a sentence, a query, etc. written in such a language will have so many words removed from it under the title of stop-words, that probably after the removal of such words the actual intended meaning of the sentence is not conveyed properly. Such a situation becomes a big challenge for NLP, AI and TM tasks including but not limited to query-processing, semantic web-development, text-analysis, text-retrieval, text-summarization, document classification and machine translation.

It is however important to be kept in mind that the language processing is an area in which development of text-processing algorithms or being able to comment on the text is relative and subject to the language and type of text including the domain of consideration.

2. Related Works

There are good numbers of research instances wherein the authors have worked on different international languages and language families[5,6] as well as Indian languages[7] in context of NLP. But there are fewer instances of research works where the authors have provided statistical analysis of more than two stop-words lists. Sadegahi and Vegas[13] have shown coincidence between Persian and English stop-words. They have defined light stop-word as a unigram stop-word containing few letters. Hence they have provided a comparison of at-least two languages. Alajmi *et al.*[1] present a statistical approach for extraction of Arabic stop-words list. They have used an approach consisting of calculation of mean probability, entropy and variance probability as well. They have used statistical analysis approach but for extraction of Arabic stop-word list and not comparative analysis of Arabic stop-word list with more than two other language stop-words lists. Similar to the research work for Arabic language, Zou *et al.*[22] proposed an automatic aggregated methodology based on statistical and information models for extraction of stop-words for Chinese language. Their results showed that their list is quite general and approach is promising enough for implementation purpose. NLP works in languages other than Arabic and Chinese have also been done. For instance, Rakholia and Saini[10,11] have worked on Gujarati language while Saini and Desai[20] have worked on Hindi language. There are also instances of research works involving multi-lingual documents[12].

There are also research instances wherein the authors have used stop-words lists for application purpose. Such an application to a specific domain like Twitter data is presented by Choy[3]. He has used Term-Frequency (TF), Inverse Document Frequency (IDF) and Term Adjacency (TA) for developing a stop words list for the Twitter data source. He has proposed a new technique using combinatorial values as an alternative measure to effectively list out stop words for Twitter data. Another work on Twitter data is presented by Saif *et al.*[14]. They have investigated the effect of removal of stop-words on the effectiveness of Twitter sentiment classification methods. They have found that stop-words do have a great effect on the studied classification methods. Kaur and Saini[9] have presented a Part-of-speech (POS) word class based categorization of Gurmukhi script Punjabi language stop-words.

Arun *et al.*[2] have explored the use of Latent Dirichlet Allocation (LDA) on stop-words and showed that it usage is employable for suitably handling the authorship attributions. They have also proved that this approach using stop-words is also effective in correct identification of author genders using stylometric studies for textual contents.

The related literature contains many instances of NLP algorithm applications wherein the application of stop-words has not been done or has been done partially too. Identification and analysis of most frequently occurring significant proper nouns in 419 Nigerian scams[16], textual analysis of digits used for designing Yahoo-group identifiers[18] and structural analysis of username segment in email addresses of MCA institutes of Gujarat state are examples of this[19].