International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6, August 2019

Root Based Stemmer for Telugu Script

Narla Swapna

Abstract: ***In this paper, a new stemmer has been proposed named as "Root based stemmer". This stemmer is strictly based on Dravidian script. Stemming can be used to pick up the effectiveness of information retrieval. In proposed Root based stemming technique, each and every token is compared against with all the words of a valid root words dictionary until a match is found. Then extract the matched string or substring from a token and identified as valid root. The present work is aimed to build dictionary based stemmer to extract valid root words for Indian languages especially for Telugu and compare the results with existing stemmers.***

Keywords: stemming, Information retrieval, Telugu

I. INTRODUCTION

Indian languages, especially the Dravidian languages (southern India languages) are rich in morphology. Complex rules existed in word formations. In the Telugu language, words are always combined with such variants, as suffix డు in రాముడు demanding the segmentation of root word రాము. Verbal variants are found to be influenced by gender such as వచ్చాడు (he has come), వచ్చింది(she has come), and with count such as వచ్చారు (they have come) that are formed with raw verb వచ్చాను(came). Telugu base word recognition is the main issue for most of the task like IR and Text categorization. "Sandhi" (Compounding) of two words transformed into a different form is a usual phenomena found in Telugu script. Raw words as document units are explored in [4, 5], adoption of morpheme model for document modeling demands for an efficient stemmer. Stemmer design is reported with maximum of 68 percent efficiency even with higher amount of computational complexity. Writing schemes engaged in Dravidian and other Indo-Aryanand scripts comprise a cross between Syllabic script methods and Phonetic script methods. Writing scheme is built on phonological values. Telugu language alphabets are formed with a character set constituting 16 vowels and 36 consonants. Akshara is the orthographic syllable of the script composed by formation of consonants and vowels following (C(C)) CV canonical construction. While a character is encoded with a single code point, a syllable (Akshara) is encoded with number of code points varying from one to nine depending on the arrangement. Decomposition of individual syllables from the word us presented in [4]. As the syllable is a meaningful entity for Indian language. In Text categorization and Information retrieval stemming is absolutely necessary.

Revised Manuscript Received on August 30, 2019.
* Correspondence Author
Dr. Narla Swapna*, Department of CSE, CMR college of Engineering & Technology Hyderabad, India

The process of stemming is, reducing the inflected or resultant words to their base words. In general, a stem is a written word form. The stem need not be the same to the morphological root of the word, it is usually enough that correlated words map to the identical stem, even if this stem is not in itself a valid root.

In Text classification, stemming tries to cut off details like exact form of a word and produce word bases as features. In 1960s, the stemming algorithms have been developed in the department of computer science.

While replicating a document, the choice of the suitable documentary unit is the first step. A documentary unit consists of set of words extracted from the sentences. The documentary unit is an approach which is widely accepted and is the best [6, 9, 13]. This is also known as word based model. A variant of word substituted with its base word (root) is one more way of representing a document, named a morpheme model. The effectiveness of the stem is the bases for measuring the performance of this method [12].

The stemmers in general are language specific. Linguistic rules are used to designing of these stemmers for a particular language. One of the substitutes for the above morpheme method is the usage of document units consisting of n-grams of characters [16]. The sequences of n character taken from the words constitute n-grams of character. The n-gram model is language independent model. The combination of the language dependent and independent models is proven to be more effective in case of European languages [11]. The paper is structured as follows ; section 2 describe the about stemming in Indian languages, section 3 explains proposed root based stemming method, section 4 illustrates results and at the last, section 5 conclusions is drawn.

II. STEMMING

The stemming technique is used to shrink the aloft of indexing and to pick up the basic IR system efficiency. It is the technique for raising IR performance is to spare searchers with ways of finding morphological variants of search items. It has been observed that the language models play a key role in identifying the valid root word/stem. Stemming technique offers two main advantages to basic IR system. First, the recall of the scheme is improved, when the query words are matched with their morphological variants in the documents.

The second benefit is it decreases the index size thereby dominating to major advantages in memory and speed. Stemmers are classified into language dependent stemmers and independent stemmers shown in figure 1.

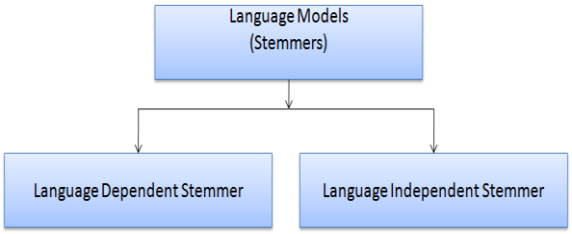

Figure 1 Classification of stemmers

A language independent model does not require any linguistic knowledge to identify the root word. N-gram is a language independent model. It is an alternative to word based models. Language dependent stemmers require linguistic knowledge to identify the root word. They are developed by using either statistical or rule based approaches. The set of rules are used by stemmer, which contains list of stems and the replacement rules to stripping of affixes. It is based on a program oriented method where the programmer has to identify every possible affixes with all replacement rules. The Porter's algorithm is an open source and commonly used approach. The major benefit of stemming algorithm is that, it is very much suitable to morphologically affluent languages like Dravidian languages. Combination of linguistic approaches with machine learning approaches results a new development in technology.

A. Stemmers for Telugu Language

Telugu is a highly inflected language. Each Telugu word is inflected in hundreds of word forms. Each inflected word starts with root and it has many suffixes [3]. The word suffix used here refers to inflections, post-positions and markers. These suffixes are affixed with each root word to generate word forms. A rule based stemmer or statistical based stemmer can be applied on Telugu. Statistical based stemmers have very less significant for the languages which have lack of corpus. Morphological tools and statistical approaches are proven to be good when the language has very less or lack of digitized dictionaries. Corpus based stemming techniques are provided by kavi Narayanamurthi et.al [10]. Three different stemming techniques are discussed in his proposal. Yet another statistical trimmer was is also proposed by Dr K.V.N Sunitha, N. Kalyani[8] for Telugu language. They use an unsupervised method for some statistical analysis to trim a Telugu word. It doesn't require any additional resource, an expert of the language.

A rule based approach for A Telugu morphological analyzer was proposed by Uma maheswararao. G [15]. The proposed methods are based on linguistic database and analyze the every word irrespective of its inflectional and derivational word. A novel method was proposed by Sunitha K.V.N, Kalyani N [7], to improve the performance of a rule based Telugu Morphological Analyzer. This is a similar to an approach which was proposed by Gaussier for suffixes identification [2]. Raw text corpus is used to acquire the suffixes. This method can also used for processing of different languages with similar behavior. A TelStem was proposed by A.P. Siva Kumar, P. Premchand, A. Govardhan[1]. It is an unsupervised Telugu stemmer using Take-all-Splits heuristic and it is not a language specific.

III. METHODOLOGY

The general workflow of Root based stemming method is demonstrated as Figure 2 and it can be separated as four sections. They are document preprocessing, document tokenization, Stopwords removal and Root based stemming.

A. Document Preprocessing:

Preprocessing is an important step in Information Retrieval and Text categorization. In this process, first a text document is read line by line from Telugu corpus and each line is pre-processed by elimination of non-Telugu characters, numerals and special characters like colons, semicolons and quotes.

B. Document Tokenization:

In this section, the given the documents, text is chopped in to chunks, this task is known as tokenization. Such chunks are called tokens. Semantically, a token is an instance of a sequence of characters that are semantically grouped into a unit for processing. Token is a general terminology of a term. In Information Retrieval and Text categorization the meaningful tokens are considered to be terms. Then a pre-processed document is tokenized and extracts the raw words. Words in Telugu text are separated by spaces and are extracted with spaces as delimiter from the document and place all raw words in *Input File*.

C. Stop words removal:

Generally, the information will have more or less frequency is detached from information in this section, because they barely contribute to the staging of retrieval. This method is called as "stop word removal". However, stop word removal is not compulsory. Removal of stop words from index collection will benefit the system by reducing index size without much effecting accuracy of the user query and availability of the item set. Elimination of word level stop words will reduce the index size. In this step, all stop words can be removed from the tokenized document and finally apply the Root based stemming model for root word identification.

D. Root based stemming

Stemming process is very essential preprocessing stage in information retrieval system. The stemming process is one of the preprocessing techniques that diminish the high dimensionality of vector space by sinking the word to its stem.

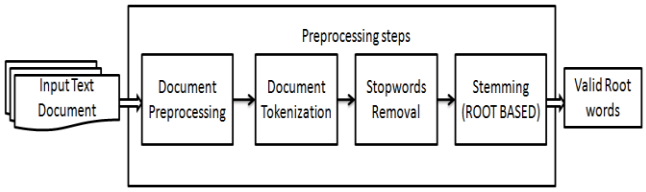

Figure 2: Workflow of Root based stemming

Stemming is mainly done by extracting any attached prefixes and suffixes from index terms before the assignment of the term. Since the stem of a term represents a vast concept than the native term, the stemming process finally raises the number of retrieved documents. In text categorization, morphological analysis (or) stemming is an absolute necessity for most complex languages. Telugu is highly inflectional language. So developing a good morphological analyzer (or) stemmer is necessary to identify the root word.

The Root based stemming method is completely dictionary based method. In this method, instead of removing prefixes and suffixes, the word extracting method is used to extract valid words from the input file based on root words dictionary. The process of root based stemming model is shown in figure 3. In Root based stemming model, one word will be choosing at a time from an *Input File*, which consisting of set of words after tokenization. That word is compared against the Telugu valid root words dictionary. If that complete word or substring matched with the dictionary, then extract that total matched string or substring and that word will be printed in a valid root file. Then next word will be selected from the *Input File* and the same process will be repeated. The same process will be continued until the *Input File* is empty.

IV. EXPERIMENT RESULTS

We conduct experiments on Telugu text Corpus, collected from online newspapers and Wikipedia. Python 3.6.0 programming is used for our work implantation. After applying root based stemming and the root words were identified, these words are validated with the help of Telugu digital dictionary was developed by University of Hyderabad (UOH).

Description about Telugu data set

In total 1169 Telugu text documents of seven different categories namely క్రీడలు(Sports), పాటలు(Songs), కథలు(Stories), సాహిత్యము(literature), వార్తలు(News), రాజకీయాలు(Politics) and నదులు(Rivers) were considered for the present study is shown in Table 1.

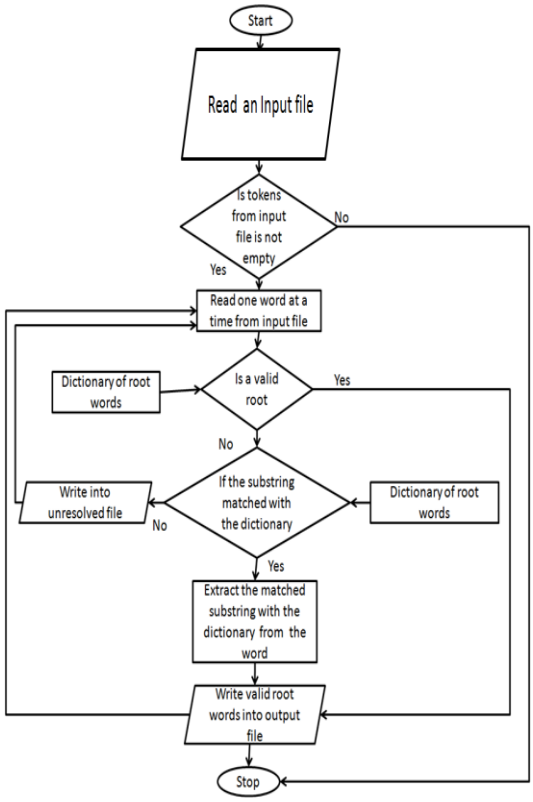

Figure 3: Process of Root based stemming model

The proposed model is applied for the same documents for which Vibhaktulu based stemming (VBS) model and suffix removal stemming (SRS) were applied. The accuracy of Root based stemming model is calculated and compared. Accuracy of the Root based stemming model is presented in Table2. It has been observed from the experiments conducted that, the Root based stemming method gives more accuracy than other language dependent existing stemming methods like VBS and SRS [14] is shown in figure 4. From the figure 4, we observed that Root based stemming model achieve 57% accuracy. The average accuracy of Root based stemming model has increased by 12.65% when compared to Suffix removal stemming.

Table 1: Telugu data set

Data set	No. of documents	No. of Words need to be identified as a root
క్రీడలు (Sports)	244	19170
పాటలు (Songs)	110	7547
కథలు (Stories)	120	57543
సాహిత్యము (Literature)	247	22551
వార్తలు (News)	100	26586
రాజకీయాలు (Politics)	268	18642
నదులు (Rivers)	80	19432
Total	**1169**	**1,71,471**

Table 2: Accuracy of Root based stemming model

Data set	Total No. of Words	No. of Words identified as stem with Root based stemming	No. of Words identified in %
క్రీడలు (Sports)	19170	9567	49.9
పాటలు (Songs)	7547	4893	64.8
కథలు (Stories)	57543	35728	62
సాహిత్యము (Literature)	22551	13549	60
వార్తలు (News)	26586	14937	56.2
రాజకీయాలు (Politics)	18642	7086	38
నదులు (Rivers)	19432	12141	62.4
Average Accuracy	**1,71,471**	**97,901**	**57**

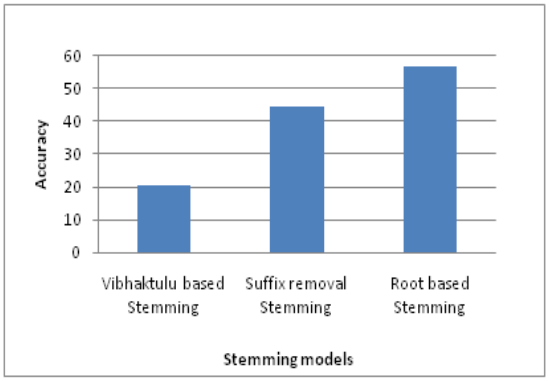

Figure 4: Comparison of Root based stemming with existing stemming models

V. CONCLUSIONS

The Root based stemming model is also well suited for different complex Indian languages like Hindi, Malayalam and Kannada. Our proposed approach will extracts valid stem words; it will become easy for stemming. In this paper, accuracy of Root based stemming model is observed and compared with other two existing stemming models. Root based stemming model accuracy is more when compared to VBS and SRS models. But, it's accuracy is very low when compared with language independent model (Pseudo syllable N-gram model).

REFERENCES

1. Damashek, M., Gauging Similarity with n-grams: Language-Independent Categorization of Text, Science, Volume 267, February 1995.
2. Eric Gaussier. Unsupervised learning of derivational morphology from inflectional lexicons. In ACL '99 Workshop Proceedings: Unsupervised Learning in Natural Language Processing , pages 24–30. ACL, 1999.
3. Ide and J. Veronis. , Introduction to the Special Issue on Word Sense Disambiguation-The State of the Art, Computational Linguistics, vol-24, issue-1, 1998, PP:2-40.
4. Indian Language Text Representation and Categorization Using Supervised Learning Algorithm M NarayanaSwamy ,M. Hanumanthappa.
5. J. Mayfield and P. McNamee, "Single N-gram stemming", Proceedings of the 26th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, 415-416, 2003.
6. Kamps J and Rijke M.D. The Effectiveness of Combining Information Retrieval Strategies for European Languages Proceedings of Symposium on Applied Computing (ACM SAC' 04),ACM Press,pp.1073-1077.
7. KVN Sunitha, N.Kalyani, "Unsupervised Stemmer to Improve Rule Based Morph Analyzer", published in the Journal - international Journal of Computer Information Systems and Industrial Management Applications, vol 2, *July 2010,* ISSN: 2150-7988, pg.nos: *179-186.*
8. KVN Sunitha, N.Kalyani, "YAST-Yet another statistical trimmer", published in the International Journal of Computer applications in Engg, Technology and Sciences Oct'2008, ISSN:0974-3596, pp:146-157.
9. M. Jenkins and D. Smith, "Conservative Stemming for Search and Indexing", In Proceedings of SIGIR'05, 2005.
10. M.Santosh Kumar, Kavi Narayan Murthy "Corpus –Based Statistical Approches for Stemming Telugu" in vishvabarath@tdil.
11. McNamee P. and Mayfield J. Character N-gram Tokenization for European language Text Retrieval Information Retrieval 7(1-2):73-97,2004.
12. Mikhel W. Berry Survey of Text Mining: Clustering Classification and Retrieval Amazon.com Springer, I Edition, September,2003.
13. Peng, F, N. Ahmed, X. Li and Y. Lu, "Context Sensitive Stemming for Web Search", Proceedings of the 30th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, 639-646, 2007.
14. Swapna Narala, B. Padmaja Rani,K. Ramakrishna, "Experiments in Telugu Language using Language Dependent and Independent Models", InternatIonal Journal of Computer Science and technology(IJCST) , Vol. 7, Issue 4, oct - Dec 2016, ISSN : 0976-8491 (online) | ISSN : 2229-4333 (print).
15. UMA MAHESWARA RAO, G. 1999. A Morphological Analyzer for Telugu (electronic form). Hyderabad: University of Hyderabad.
16. Yang Y. An evaluation of statistical approaches to text categorization Information Retrieval 1,69-90,1999.

AUTHORS PROFILE

Dr. Narla Swapna, working as an Associate professor in the Dept. of CSE, CMR College of Engineering & Technology, Hyderabad. She obtained Ph.D from Jawaharlal Nehru Technological University, Hyderabad and M.Tech from Jawaharlal Nehru Technological University, Anathapur. She has 12 years of teaching experience and 15 publications in various National, International Conference and Journals. She held different positions like JKC core team member for nine years for 2006-2015, single point of contact (SPOC), NBA coordinator, TPO and in charge HOD. Her research interests include Machine Learning, Information Retrieval system, Data mining and Natural Language Processing. She is also a Member of various Technical Associations including ISTE, IEEE etc.