Literature Review: Stemming Algorithms for Indian and Non-Indian Languages

[I]R. Vijaya Lakshmi, [II]Dr. S. Britto Ramesh Kumar
[I]St Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India
[II]Research Advisor, St Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India

Abstract

Stemming is a technique used for reducing inflected words to their stem or root form. This is applicable for both the suffix as well as prefix. Stemming is a preprocessing step in text mining application and commonly used for Natural Language Processing (NLP). A stemmer can execute operation of altering morphologically identical words to root word without performing morphological analysis of that term. It is useful in many areas of computational linguistics and information retrieval work for refining their performance. The purpose of stemming is used to progress the effectiveness of information retrieval. The goal of stemming is to diminish the inflectional forms and sometimes derivationally related forms of a word to a common root form. In this paper we have discussed about proposed method, Data set, Accuracy and Approaches used in recently developed stemming algorithms for Indian and Non-Indian languages.

Keywords

Natural Language Processing, Morphological analysis, Stemming, Emille, Information Retrieval, Na'ive Bayesian, F-measure, WordNet

I. Introduction

Stemming plays an important role in Information Retrieval System (IRS) for improving the performance of all languages. The goal of stemming is to diminish inflectional and derivational variant forms of a word to a common base form. A stemmer can execute operation of transforming morphologically identical words to root word without performing morphological analysis of that term. This makes stemming an attractive preference to raise the ability of matching query and document vocabulary in information retrieval system. Many natural languages like Dravidian languages (Tamil, Telugu, Malayalam and Kannada), Indo

Aryan languages (Hindi, Bengali, Marathi, Gujarati) searching quality is increased because of using stemming algorithm. Thus various stemming are developed for various languages, but each one has its own advantages as well as limitations. Most of the stemming algorithms are language dependent. So there is an urgent need to develop language independent stemming algorithm to increase the searching efficiency.

II. Related Work

The different stemming for both Indian and non -Indian language accuracy and errors was found by Bijal et.al [1]. A new improved light stemming algorithm proposed by Thangarasu et.al for less computational steps which is used to get good stemmed Tamil words. Also uses K-means clustering for the performance of Tamil language [3, 10]. A new stemmer "maulik" was proposed for Hindi language by Mishra et.al [5] using Devanagiri script and hybrid approach. Anjali Ganesh Jivani has discussed various methods of stemming and their comparisons in terms of usage, advantages as well as limitations [6]. The fundamental difference between stemming and lemmatization is also discussed. Perhaps developing good lemmatizer could help in achieving the goal. But this paper does not deal with recently developed stemming algorithms. A new light stemming technique was introduced and compared this with other stemmers to show the improvement of search effectiveness in Arabic language by Mohamad et.al [7]. An approach for finding out the stems from the text in Bengali was presented by Das et.al [8]. In this paper, they maintained two different hash tables, first one containing all possible nominal inflections and the second one containing all possible verbal inflections for Bengali language. An unsupervised approach for the development of stemmer in Urdu and Marathi language had been presented by Husain [9]. And frequency based and length based stripping are proposed for rule generation. But author was using N gram method. N gram approach requires large memory size. In this research the approach is based on different stemming algorithms for different language. Stemming and Lemmatization: A Comparison of Retrieval proposes to compare document retrieval precision performances based on language modeling techniques, particularly stemming and lemmatization [11]. A Comprehensive Study on Stemming Algorithms by Kasthuri et.al focuses different stemming methods and their comparisons in terms of usage, pros and cons of them [15]. But this paper does deal with recently developed stemmer for Indian languages.

III. Contextual To Stemming

Stemming algorithms try to build the basic forms of words, i.e. strip the plural's' from nouns, the 'ing' from verbs. A stem is a natural group of words with equal or much related meaning. After this process, every word is represented by its stem.

IV. Errors in Stemming

There are mainly two errors in stemming over stemming and under stemming. Over stemming is defined as when two words with different stems are stemmed to the same root. This is also known as a false positive. Under stemming is defined as when two words that should be stemmed to the same root are not. This is also known as a false negative.

V. Stemming Algorithms for Indian Languages

A. Bengali Stemming

There are various stemming algorithms proposed for Bengali language. Bengali language is the morphologically rich in nature. Recently developed Bengali stemming algorithm has to deal with nominal and pronominal inflections and verbal inflections in Bengali. Two different hash tables were maintained for this process and all possible nominal inflections are collected in the first inflectional suffixes and all possible verbal inflections are collected in second inflectional suffixes for Bengali language. Now from the input words, locate the suffix part of maximum length of the suffix part in input words are taken and finding the predefined suffixes which are matched to the input word. If the input word is not matched means then it is treated as root word. In every stemming, the rule based approach of the suffix stripping part is ended using some predefined rules of that stemming technique. Only one root word is undertaken from every inflectional word for this method. Proposed approach gives the MAP value of 0.4748, which is quite good compared to the previous works for unsupervised stemming process.

B. Urdu and Marathi Stemming

Both Urdu and Marathi languages are highly inflected in nature. Recently developed language independent stemmer deals with Frequency based stripping and length based stripping for generating suffix rules. This proposed approach is based on n-gram splitting model. For Urdu and Marathi language there are two approach used. First is the Length based approach which is very simple suffix stripping approach. The second approach is Frequency based. For evaluation1200 words are extracted from the EMILLE corpus. But N gram approach requires large memory space.

C. Kokborok Stemming

Several stemmers have been developed for a large number of languages including Indian languages; however not much work has been done in Kokborok, a native language of Tripura. For Kokborok language, a simple rule based stemmer was considered and it uses an affix stripping algorithm. This algorithm is used to measure the reduction of inflected words to the stem or root form. Then affix stripping algorithm was developed for reducing joined Kokborok words to its stem or root. Nearly 32578 documents are collected from corpus and it provided the result based on minimum suffix stripping algorithm and for maximum suffix stripping algorithm. Without having strong linguistic background we could not establish this kind of stemming algorithms.

D. Hindi Stemming

Regarding Indian languages, extremely less work has been discussed for performing stemming algorithms. But much of work has done for stemming in English and other European languages. Hindi is the national language in India. But not much linguistic assets are available for Hindi as research is still at early stage for Hindi. This paper discusses Hindi stemmer for nouns [2].

Table 1: Stemming Algorithms for Indian languages

Year	Tested on Language	Total Words	Name of Authors	Proposed Method	Approach Used	Dataset	Accuracy
2010	Punjabi	28000	Dinesh Kumar, Prince Rana	Brute Force Technique	Suffix Stripping	Pardeep Punjabi to English Dictionary	80.73%
2011	Bengali	123047	Suprabhat Das, Pabitra Mitra	Method Proposed by Porter	Suffix Stripping	Bengali collection of the FIRE 2010 Data Set	96.27%
2012	Gujarati	3000	Juhi Ameta, Nisheeth Joshi, Iti Mathur	Longest matched	Rule based	EMILLE corpus	91.5%
2012	Urdu	1200	Shahid Husain	n-gram stripping model	1)Length based 2)Frequency based	EMILLE Corpus	1)84.27%, 2)79.63%
2012	Kokborok	32578	Braja Gopal Patra, Dipankar Das	Rule Based Stemmer	1)Minimum Suffix Stripping 2) Maximum Suffix Stripping	Corpus from Story books and Holy Bible	1)80.02% 2)85.13%
2012	Hindi	15000	Upendra Mishra, Chandra Prakash	Brute force technique	Suffix stripping.	Not mentioned	91.59%
2012	Marathi	1200	Shahid Husain	n-gram stripping model	1)Length based 2) Frequency based	Marathi Corpus	1)63.5%, 2)82.68%
2013	Tamil	7000	Thangarsu, Manavalan	Light Stemmer	Light stemming	Online Tamil documents	99%
2014	Hindi	100 documents	Vishal Gupta	Suffix stripping	Rule based suffix stripping	Documents from Hindi Newspaper	83.65%
2014	English & Tamil	900	Kasthuri, Britto Ramesh Kumar	Dynamic Programming Technique	Phonetic based stem generation	EMILLE Corpus	99%

This stemmer applied to suffix stripping rule based approach for performing stemming of nouns. There are three techniques used for designing the suffix rules such as statistical stripping, rule based approach and stripping of suffixes. Noun plays vital role in this Hindi stemmer and this stemmer applies suffix stripping rule based approach for performing stemming of nouns. After analyzing news articles from popular Hindi news papers, corresponding stemming rules were developed for each suffix and 16 noun suffixes have been generated. The accuracy of this stemmer can be improved by adding more suffixes of Hindi nouns and more stemming rules.

E. Tamil Stemming

Stemmer provides to decrease the size of the index files in the IRS. This is particularly true in case of a morphologically rich language like Tamil, where a single word may take many forms. The aim is to make sure that related words map to common stem.

There are several stemmer algorithms have been proposed and evaluated for various Indian languages such as Hindi, Marathi, Bengali, Urdu, Malayalam, etc. Recently developed stemmer for Tamil using K-mean clustering approach is used to get stemmed Tamil word with less computational stages. In order to improve the performance of Tamil language the stemmed Tamil Words are clustered by using K-mean algorithm. Improved light stemming is robust. Improved light stemmer algorithm is fitting for IRS of Tamil language. Since it execute efficiently for morphological rich language Tamil.

A Rule Based Iterative Affix Stripping Stemming Algorithm for Tamil discusses about the implementation of a stemming algorithm that is accessible as Open Source Software. This algorithm was implemented using rule based iterative affix stripping algorithm [13]. But this is the language dependent approach. Without having strong linguistic background we could not create this kind of stemmer.

Monolingual phonetic based stem is an innovative attempt is being made to develop an algorithm for novel conflation method that exploits the phonetic quality of words and uses some standard Natural Language Processing tools like Levenshtein Distance and Longest Common Subsequence for Stemming process [14]. The evaluation has been made on 900 English words and 900 Tamil words extracted from Emille Corpus. Actual application of the designed algorithm to the testing data gave nearly 100% results in terms of producing correct stems. This approach is used to find stem word of English as well as Tamil Languages. Even this approach can support Indian and Non-Indian languages. Table 1 shows that Accuracy rate, proposed approach, Data set details about Stemming algorithms for Indian Languages.

VI. Stemming Algorithms for Non-Indian Languages

1. Persian Stemming

Structural Rule-based stemming for Persian language uses the structure of words and morphological rules of the language for recognizing the stem of each word. For this 33 rules are collected for describing a structural rule-based stemmer and developed a rule-based stemmer for Persian language. For this purpose heuristic rules based on Persian word structures are used. To improve the stemmer by focusing on the rules which are defined to recognize the stem, may be able to increase the precision.

2. Arabic Stemming

Text preprocessing of Arabic Language is a challenge and critical stage in Text Categorization particularly and Text Mining generally. But now days there are various stemming algorithm were proposed. An efficient hybrid method is proposed for stemming Arabic text. The effectiveness of the four methods was evaluated and compared in term of the accuracy of the Na'ive Bayesian classifier. The obtained results illustrate that using the proposed stemmer enhances the performance of Arabic Text Categorization.

Stemming Effectiveness in Clustering of Arabic Documents proposes stemming algorithm for Arabic Language. Arabic language is more complex than most other languages. Evaluation can be carried out in three stemming techniques: root-based Stemming, without stemming and light Stemming. One of the famous and generally used clustering algorithm is K-mean is applied for clustering. They made an evaluation which is depends on recall, precision and F-measure methods. From experiments, results show that light stemming achieved best results in terms of recall, precision and F-measure. Yet no a complete stemmer for this language is available: The existing stemmers not have a high performance.

3. Turkish Stemming

Turkish is an agglutinative language with a highly productive inflectional and derivational morphology. We learn information retrieval (IR) on Turkish texts using a large-scale test collection that contains 408,305 documents and 72 ad hoc queries. We observe the effects of several stemming options and query-document matching functions on retrieval routine. Information retrieval for Turkish language paper shows that a simple word truncation approach. Word truncation approach uses language dependent corpus.

Performance Analysis and Improvement of Turkish Broadcast News Retrieval paper deals with the retrieval of spoken information in Turkish language [4]. Conventional speech retrieval systems perform indexing and retrieval over automatic speech recognition transcripts, which contain errors either because of out-of-vocabulary words or ASR inexactness. They use sub word units as recognition and indexing units to diminish the OOV rate and index alternative recognition hypotheses to handle ASR errors. Evaluated on our Turkish Broadcast News Corpus with two types of speech retrieval systems: spoken term detection (STD) and a spoken document retrieval (SDR) system. To evaluate the SDR system, they also build a spoken information retrieval (IR) collection. Experiments showed that word segmentation algorithms are quite useful for both tasks. Still there are some limitations available in the existing Turkish stemming algorithms. Table 2 shows that Accuracy rate, proposed approach, Data set details about Stemming algorithms for Non-Indian Languages.

VII. Conclusion

Stemming plays a dynamic role in information retrieval system and its effect is very huge, related to that analysis on various stemming algorithms. In this paper we have studied about various stemming algorithms and its efficiency on various Indian and Non-Indian languages. Stemming algorithm is also useful in dropping the size of index files as the number of words to be indexed are reduced to common forms or so called stems. Most of stemming algorithms are based on rule based approach. The enactment of rule based stemmer is superior to some well-known method like brute force technique. This is not quite sufficient for information retrieval system. Some stemming algorithms reaches better in some area, other reaches better in some other area. So that in

Table 2: Stemming Algorithms for Non-Indian languages

Year	Tested on Language	Total Words	Name of Authors	Proposed Method	Approach Used	Dataset	Accuracy
2010	Persian	166477	Elaheh Rahimtoroghi, Hesham Faili, Azadeh Shakery	Rule Based stemmer	Structural approach and Morphological rules	Newspaper articles	The precision has been increased by 4.83%.
2011	Indonesian	Not mentioned	Ayu Purwarianti	Non-deterministic Finite Automata	Suffix Stripping	Dictionary	Not mentioned
2012	Arabic	11347	Mohamad Ababneh, Riyad Al-Shalabi	Rule Based light stemmer	Root extraction stemmer and light stemmer	Not mentioned	Not mentioned
2012	Arabic	4763	Osama A. Ghanem, Wesam M. Ashour	K-means Algorithm	Clustering	BBC Arabic Corpus	Not mentioned
2012	Turkish	3697	Sidikka Parlak, Murat Saraclar	Length based	Rule Based	Broadcast News Corpus	Not mentioned

future, researchers will go for more number of executions for the stemming algorithm methods and their benefits for various Indian and Non-Indian languages information retrieval system. There is an urgent need to develop language independent stemmer for all languages. In future we try to create the new stemming approach to find out stem word for many languages.

References

[1] *Dalwadi Bijal, Suthar Sanket, "Overview of Stemming Algorithm for Indian and Non-Indian Languages", International Journal of Computer Sciences and Information Technologies (IJCSIT) Vol. 5 (2), PP. 1144-1146, 2014.*

[2] *Vishal Gupta, "Hindi Rule Based Stemmer for Nouns", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1, January 2014.*

[3] *M.Thangarasu. R.Manavalan, "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, July 2013.*

[4] *Siddika Parlak, Murat Saraclar, "Performance Analysis and Improvement of Turkish Broadcast News Retrieval", IEEE Transactions and audio, Speech and Language Processing, Vol. 20, No. 3, PP 731-740 March 2012.*

[5] *Upendra Mishra, Chandra Prakash, "MAULIK: An Effective Stemmer for Hindi Language", International Journal on Computer Science and Engineering (IJCSE) Vol. 4 No. 5, PP.711-717, May 2012*

[6] *Ms. Anjali Ganesh Jivani, "A Comparative Study of Stemming Algorithms", International Journal of Computer Technology and Applications, Vol.2 (6), PP 1930-1938, NOV-DEC 2011.*

[7] *Mohamad Ababneh, Riyad Al-Shalabi, Ghassan Kanaan, Alaa Al-Nobani, "Building an Effective Rule-Based Light Stemmer for Arabic Language to Improve Search Effectiveness", The International Arab Journal of Information Technology, Vol. 9, No. 4, PP.368-372, July 2012.*

[8] *Suprabhat Das, Pabitra Mitra, "A Rule-based Approach of Stemming for Inflectional and Derivational Words in Bengali", Proceeding of the IEEE Students' Technology Symposium, PP.14-16, January, 2011.*

[9] *Mohd. Shahid Husain, "AN UNSUPERVISED APPROACH TO DEVELOP STEMMER", International Journal on Natural Language Computing (IJNLC) Vol. 1, No.2, August 2012.*

[10] *M.Thangarasu., R.Manavalan, "A Literature Review: Stemming Algorithms for Indian Languages", International Journal of Computer Trends and Technology (IJCTT), volume 4 Issue 8, August 2013.*

[11] *Vimala Balakrishnan, Ethel Lloyd-Yemoh, "Stemming and Lemmatization: A Comparison of Retrieval Performances", Lecture Notes on Software Engineering, Vol. 2, No. 3, August 2014.*

[12] *Dhamodharan Rajalingam, "A Rule Based Iterative Affix Stripping Stemming Algorithm for Tamil", vol 132, PP-583-590, 2012.*

[13] *M.Kasthuri, S.Britto Ramesh Kumar, "Multilingual Phonetic Based Stem Generation", Second International Conference on Emerging Research in computing, Information, Communication and Applications(ERCICA), pp 243-248, August 2014.*

[14] *M.Kasthuri, S. Britto Ramesh Kumar, "A Comprehensive Study on Stemming Algorithms", Software Quality Engineering for Enterprise and Cloud Computing (5th International Conference on Semantic E-Business and Enterprise Computing), ISBN: 978-93-82062-64-6, P.No:33-40, 13-15 December 2012.*