A Literature Survey: Stemming Algorithm for Odia Language

[1]Dhabal Prasad Sethi, [2] Sanjit Kumar Barik

ABSTRACT

Stemming is the process of conflating morphological variants to a common stem or root. For better information retrieval, stemming algorithm is used. Suppose the user who does not know the exact topic to retrieve but he know some keyword then after typing some word it may fetches the exact topic along with all the related form of that topic. Using stemmer user can get better result. So stemming is generally called as recall-enhancing device. In this article we study the different algorithm used in Odia language for stemming.

Keywords: Stemmer, Precision, Recall, Informational Retrieval, Indexing, Survey

I.INTRODUCTION

Stemming is the process of removing the inflectional or derivational suffixes from words to stem or root. For information retrieval system stemming plays important role. Before there is a controversy about root or stems which one is best as terms for indexing in IR system? Some researches before said that root based technique gives better performance than stem based for indexing purpose. Later some researchers' gives the feedback that stem based technique gives better performance than root based. So the beginner researcher will get confused? Before the researchers experimented a small collection of text so that they said root based technique is better for indexing. Later the researches make the experiment and takes large number of text for testing and found stem based indexing retrieves more data. The recent research concludes that stem based indexing used in information retrieval is best. There are two common terms named precision and recalls are used in IR. What is precision? And what is recall? The solution is: Precision is the fraction of the documents retrieved that are relevant to the user's information need. Precision=| {relevant documents} intersection {retrieved documents}|/| {retrieved documents} |.Recall is the fraction of the documents that are relevant to the query that are successfully retrieved. Recall=| {non-relevant documents} intersection {retrieved documents}|/| {non-relevant documents}|.

This paper is organized as II. describes the literature survey III. describes different algorithms in stemming IV. describes common error in stemming V. describes applications of stemmer VI presents the conclusion.

First Author Name*: Dhabal Prasad Sethi, Lecturer in Computer Science & Engineering, Government College of Engineering,Keonjhar,Odisha.*

Second Author Name*: Sanjit Kumar Barik, Lecturer in Computer Science and Engineering, Government College of Engineering, Keonjhar,Odisha.*

II.LITERATURE SURVEY

Sampa et.l [1] presented a paper named a suffix stripping algorithm for Odia stemmer. She uses the suffix stripping algorithm to remove the inflectional (bibhakti) suffixes. That algorithm predicts 88% result. Lastly she draws a diagram of stemmer using finite automata.

R.C Balbantray, B.Sahoo, M.Swain, D.K, Sahoo [2] presented a paper IIT-Bh FIRE2012 Submission: MET Track Odia. They have used the affix removal method. They firstly store the root word in a dictionary, a stop word list in another dictionary. Their system read a folder containing text files as input. When the input files matched the stop word dictionary it removes the stop word then matched with the root dictionary if found they mentioned no further processing required, they get the root .If does not match the dictionary then goes to further processing. When the input matched the suffix dictionary then removes the suffix and no further processing required.

R.C Balabantaray,B Sahoo, D.K Sahoo, M Swain [3] presented a paper Odia Text Summarization using Stemmer. They summarize the Odia paragraph using stemming algorithm. So text summarization is one of the applications of stemmer.

Dhabal Prasad Sethi [4] presented a paper named design of lightweight stemmer for Odia derivational suffixes. He mentions the technique which recursively removes the suffix. First he has tested the derivational suffixes using suffix stripping algorithm and he found the result 66.25% .Because some words are over-stemming and some words are under stemming .To solve that over-stemming problem he design an algorithm that solves the over stemming problem .This algorithm predicts 85% result approximately.

III.DIFFERENT ALGORITHMS IN STEMMING

1) Brute Force Algorithm: Brute force is the straight forward algorithm. This algorithm employee two table, one table is root word and another table is inflection word. To stem a word the table is queried to find a matching inflection, if found the root form is returned. Example if user enter the word "dogs" as input, it searches for the word "dog" in the list. When match found, it display the result.

2) Suffix Stripping Algorithm: Suffix Stripping algorithm is an approach which removes the suffixes, if a word ends with a certain character sequence. It is small and efficient algorithm and does not hamper the linguistic claims. This algorithm is developed by martin porter in 1980.Now this algorithm is widely used in different language.

3) Suffix Substitution: This algorithm is the improved version of suffix stripping algorithm. In this algorithm, instead of removing the suffixes, it substitutes another suffix**.**

ISSN: 2278 – 1323

4) Affix Removal Algorithm: This algorithm removes the affixes e.g. prefixes, suffixes. It removes the longest possible string of character from a word using set of rules.
5) Matching Algorithm: This algorithm uses a stem database. To stem a word this algorithm tries to match in the stem dictionary. Example the prefix "be" which is the stem form of "be"," been", "being". If user enters the word "besides" it stems to "be".
6) Stochastic algorithm: Stochastic algorithm uses the probability to identify the root form of a word. This algorithm is trained on a table of root form to inflected form to develop the probabilistic model. This model follows the complex linguistic rules, similar to suffix stripping. Stemming is done by inputting the inflected form to the trained model and the model produce the root from according the internal rule set.
7) Hybrid approach: This algorithm combines more than two approaches. It may be combine the rule based method along with statistic method.
8) **N-Gram**: Many stemming techniques use the n-gram approach of a word to choose the correct stem for a word.

IV.COMMON ERROR IN STEMMING

Generally there are two common errors are done in stemming .They are over-stemming, under-stemming .Over stemming means when words are not morphological variants are conflated. Example in English language the words "wander" and "wand" are stemmed to "wand". The error is that ending 'er' of 'wander' is considered a suffix whereas it is the actual stem word. Under-stemming means those words are morphological variants are not conflated. Example compile is stemmed to comp and compiling to compil [10].

V.APPLICATIONS OF STEMMER

1) Morphology: Morphology is the study of internal structure of word. Morpheme is the smallest individual unit of a word. Suppose the word PILAMANE, the morphemes are PILA and MANE.
2) Information Retrieval: Information retrieval is the process of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or full text indexing. Automatic information retrieval systems are used to reduce what has been called "information overload" .Web search engines are the most visible IR applications. Other applications of information retrieval are digital libraries, information filtering, recommender systems, media search (blog search, image retrieval, music retrieval, news search, speech retrieval, video retrieval), search engine (desktop search, enterprise search, mobile search, social search, web search) [9].
3) Auto Text Summarization: It is the process of reducing a text document using a computer program in order to create a summary that keeps the most important points of the original documents.
4) Text Classification/Document Classification: Document classification is the process of assigning/classifying documents to one or more classes or category. The text documents to be classified which may be texts, images, music etc.
5) Machine Translation: Machine translation is sub-branch of computational linguistics that investigates the use of software to translate text or speech from one natural language to another .Machine translation substitutes words in one language for words in another [5].
6) Indexing: Indexing is a term used in information retrieval to collect, parses, and stores data to facilitate fast and accurate retrieve.
7) Question Answering System: Question answering system is a field of information retrieval and natural language processing (NLP), which is concerned with building computer systems that automatically answer the question which produced by humans in a natural language.
8) Cross-Language Information Retrieval (CLIR): It is a subfield of information retrieval dealing with information retrieving written in a language different from the language of the user's query. Example suppose the user enter the query in English, it retrieve relevant document written in Odia.
9) POS Tagging: A part-of-speech tagging is the process of classifying text documents of any language to its respective part-of-speech category e.g. noun, pronoun, verb, adjective, adverb etc.

VI.CONCLUSION

In this article we studied the different algorithm used by different author for stemmer in Odia and know the basic concepts of stemmer, the confusion between stem or root based algorithm which one is better for information retrieval and lastly its applications. It is the most important tool used in language software development.

ACKNOWLEDGEMENT

We would like to thanks to one of our colleague Mr. Debasish Mohanta of dept. electronics who has given his valuable suggestion.

REFERENCES

[1]A suffix stripping algorithm for odia stemmer by samapa chaupatnaik,sohag sunder nanda,sanghamitra mohanty at international journal of computational linguistic and natural language processing volume1
[2]R.C Balbantray, B.Sahoo ,M.Swain, D.K,Sahoo presented a paper IIT-Bh FIRE2012 Submission: MET Track odia.
[3]Odia Text Summarization using Stemmer presented by R.C Balabantaray, B Sahoo,D.K Sahoo, M swain from CLIA Lab, IIIT, Bhubaneswar, Odisha at International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 1– No.3, February 2012 – www.ijais.org
[4]Design of lightweight stemmer for odia derivational suffixes by Dhabal Prasad Sethi, Government College of Engineering, Keonjhar, Odisha at International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 12, December 2013 page 4594-4597
[5]en.wikipedia.org/wiki/machine taransalation
[6]en.wikipedia.org/wiki/question_answering
[7]en.wikipedia.org/wiki/cross-language-information_retrieval
[8]en.wikipedia.org/wiki/automatic-summarization
[9]en.wikipedia.org/wiki/information_retrieval
[10]A light stemmer for Hindi by Ananthakrishnan Ramanathan, Durgesh D Rao from NCST, Navi Mumbai.

BIOGRAPHY

Dhabal Prasad Sethi is working as a lecturer in CSE at Government College of Engineering, Keonjhar, Odisha. He has completed his Bachelor of Engineering from BIET, Bhadrak in 2006 and then completed his Master of Engineering with specialization in Knowledge Engineering from Post Graduate Department of Computer Science & Application, Utkal University, Bhubaneswar, Odisha in 2011.He has presented 3nos. of paper at international journals. This is his 4th no at the international level. His research area of interest is natural language processing, information retrieval, data mining and software engineering.

Sanjit Kumar Barik is working as a Lecturer in Computer Science & Engineering at Government college of Engineering, Keonjhar, Odisha. He has completed his MCA from CET, BBSR and then completed his MTECH Computer Science& Engineering from NIT, Rourkela. He has more than 6years experience in teaching. His research area of interest is data structure and distributed system.

