IIIT-Bh FIRE 2012 Submission: MET Track Odia

R.C. Balabantaray
IIIT Bhubaneswar
rakeshbray@gmail.com

B. Sahoo
IIIT Bhubaneswar
s.bibhuprasad@gmail.com

M. Swain
IIIT Bhubaneswar
monalisaswain1988@gmail.com

D.K. Sahoo
IIIT Bhubaneswar
deepsahoo@gmail.com

ABSTRACT

Stemmers attempt to reduce a word to its stem or root form and are used widely in information retrieval tasks to increase the recall rate. There is no stemming algorithm has been standardized for Indo-Aryan languages. The need for good stemming algorithms for these languages has increased in the wake of search and retrieval system. A standard stemmer of each specific language is a basic requirement for a multilingual search engine. Though Odia language is a part of this, a standard Odia Stemmer should be a part of the multilingual search engine. This paper is a little perseverance towards the development of an Odia Stemmer in the field of Information Retrieval which is based on affix removal method.

Keywords: Information Retrieval, Search Engine, Stemming, Indexing, Stop words.

1. INTRODUCTION

Automatic Information Retrieval is the process that extracts the information related to the specific query supplied by the user. The queries supplied by the user composed of a set of words which are interrelated by the Boolean operators like AND, OR. The System responds by identifying the documents containing the combination of the query words. The basic idea about the search engine is based on the indexing the content of the documents and then retrieve the documents by matching the query term with the indexed content.

Stemmer is one of many tools used in information retrieval to combat the vocabulary mismatch problem, in which query words do not match document words. Number of stemmers has been developed for different languages, which takes on to reduce a word to its root form, thereby reducing the size of index. The stem, in information retrieval purpose need not be a valid root of the word; it maps all the similar or related words to the same root. It might be possible that (computation; compute; computing) are reduced to the same root word "comput" rather than base word "compute". Stemming doesn't deals with the linguistic correctness of the root word, however algorithms which also deal with the linguistic correctness are known as lemmatiser, which is out of the scope of this paper. The retrieval effectiveness is evaluated by two important measures i.e. recall and precession. Recall is the relevant documents actually retrieved where as the precession is the proportion of the retrieved documents that is found to be relevant to the user's information need. In Natural Language Processing (NLP), stemmer and lemmatizer are used to extract the lexical and grammatical morphemes of the original words. In multilingual search language specific stemmers are required in multiple languages. In this paper we have given an idea towards the development of an Odia Stemmer which can be a part of multilingual search and can play an active role to enhance the performance of the search engine.

2. MOTIVATION TOWARDS ODIA LANGUAGE

Odia language which has a history of more than one thousand years is one of the most familiar branches of Indo-Aryan subfamily of languages. According to the 8th schedule of Indian constitution there are 22 major languages of India which are recognized by Indian constitution. Odia (Officially changed from Oriya to Odia from November 2011) is one of them which is mainly spoken by the people live in the state Odisha (Officially changed from Orissa to Odisha from November 2011). It is belongs to the Indo-Aryan language of the Indo-European language family. It is the official language of Odisha and the second official language of Jharkhand. Though it is an official Indian language, and it is spoken by 31 million people (83.33% of total population of Odisha) in Odisha, and some other neighbor states like West Bengal, Jharkhand, Chhattisgarh and Andhra Pradesh. There are also significant number of Odia speaking people in main cities of India i.e. Kolkata, Mumbai, Delhi, Chennai, Bangalore, Hyderabad, Pune, Pondicherry, Gurgaon etc. Odia is not only spoken in India. Odia people are also in some of the prominent countries of the World e.g. USA, UK, Canada, UAE, SriLanka, Singapore, Malaysia, Burma, Indonesia. There are 45 million Odia speaking people living in globally. So we are motivated towards Odia language and working for this language.

3. METHODOLOGY

Out of different stemming algorithm our work includes affix removal method. In this method, basically the valid affixes (suffix and prefix) are removed in order to get the root word. In our work we have maintained a dictionary containing root words. A stop word list also used which contains the Odia stop words (around 216 stop words). Our system takes a folder containing one or more text files as input and makes the content of the files into individual tokens. Then all the stop words are removed from the input file, because the stop words are not required for the searching purpose. After the removal of stop words, for each token, first it go to the dictionary to check whether the word itself a root word or not. If it matches with the dictionary then there is no need of further processing. It directly returns the root word. If it does not match with the dictionary then it goes for further processing. Here first the suffix is removed from the token. We have considered the longest suffix for removal here. Then it again crosschecks with the dictionary after the removal of suffix. If it matches, then there is no need of further processing. Otherwise it goes for further processing. This time the prefix is removed from the string and gives the required root.

In case of Odia language, the verb form consists of maximum elements "verb root + aspect + auxiliary + tens+ agreement" and minimum elements "verb root + agreement". In our system we handle all the three tenses and four aspects i.e. in total 12 aspects of the verbal form including some hypothetical tenses. In case of noun, our system includes plural suffixes, honorific suffixes, and some adjectival suffixes. Some exceptional case also handled in our system, like **suppletion[5]**. Suppletion is a morphological process of word formation where the inflected form does not show any relation with the root form. For example, the root of "went" is "go", root of "best" is "good". In Odia this type of exception is also there. i.e. the root of "ଗଲା" (gala-went) is "ଯା" (jaa-go).

4. RESULT AND DISCUSSION

Our stemmer is evaluated with an open source search engine Terrier-3.5. For testing purpose we have taken 50 queries and the manual judgment file for Odia language which have already submitted to FIRE-2012. We have evaluated the retrieval result both for baseline result and the result using our stemmer. It shows a little improvement in the evaluation result i.e. in average precession rate. The evaluation result is as follows:

Table: 1 Evaluation result for both baseline and with stemmer in Terrier-3.5

Evaluation	Baseline	With stemmer
No of Retrieved Documents	15896	16235
No of Relevant Documents	90	90
No of Relevant Retrieved Documents	86	85
Average Precession	0.3441	0.3866

There are some common issues in our Stemmer. We have faced difficulties to solve some problems listed below.

1. There are some non-finite verbs like "ପଢ଼" (padha-read) having no vowel addition makes problem to handle.

   Ex: "ପଢ଼" (padha-read), "ବସ" (basa-sit), "ହସ" (hasa-laugh).

2. The words having same suffix but different root.

   Ex: ମ ିତ (mandita) → ମ ନ (manda***na***)

   ଅବଧାରିତ (abadharita)→ ଅବଧାରଣ (abadhara***Na)***

   ଉ ସିତ (ullasita)→ ଉ ାସ (ull***A***sa)

   କବଳିତ(kabalita) → କବଳ (kabala)

There is confusion between the noun and verb stem

Ex: ଦେ   (dekhi)→ ଦେ   (dekh)

ଉଠି (uThi)→ ଉ   (uTh)

ସକାଳେ (sakale)→ ସକାଳ (sakala) (ସକା)

ସକାଳୁ (sakalu)→ ସକାଳ (sakala)(ସକା)

Here at a time one can give the correct stem either noun or verb.

5. CONCLUSION

The stemmer is evaluated with an open source search engine Terrier-3.5 with 50 queries and the manual judgment file for Odia language given to the FIRE-2012. The average precession rate has little bit improved in the evaluation result with the stemmer as compare to baseline evaluation result. Now we are working upon it for further improvement. Normally for a word one can analyze different things like its category, tens, aspect, gender, case marker etc. We are working upon it to incorporate these things in our stemmer, in order to improve the efficiency for various kinds of applications along with the information retrieval.

6. ACKNOWLEDGEMENT

We are very much thankful to DIT, Govt. of India for the financial help for this work.

7. REFFERENCES

[1] M. F. Porter., "An algorithm for suffix stripping. Program: electronic library & information systems," 14(3):130-137, July 1980.

[2] J. Xu and W. B. Croft., "Corpus-based stemming using co-occurrence of word variants.", ACM Trans. Inf. Syst., 16(1):61-81, 1998.

[3] M. Z. Islam, M. N. Uddin, and M. Khan, "A light weight stemmer for bengali and its use in spelling checker", In Proc. of 1st International Conference on Digital Communications and Computer Applications (DCCA2007), 2007.

[4] Payas Gupta, "Stemmer for Indo-Aryan Languages: A Survey",www.mysmu.edu/phdis2008/payas.gupta.2008/papers/IS701.pdf

[5] S.K.Lenka, "Morphosyntax; A study of Oriya", on published Phd Dissertation, Banaras Hindu University, PP. 36-79, 2011.