Automatic Text Summarization for Oriya Language

Sitanath Biswas
Indus College of Engineering, Bhubaneswar

Sweta Acharya
College of Engineering & Technology, Bhubaneswar

Sujata Dash, PhD
North Odisha University, Baripada

ABSTRACT

With the coming of the information revolution, electronic documents are becoming a principle media of business and academic information. Thousands and thousands of electronic documents are produced and made available on the internet each day. In order to fully utilizing these on-line documents effectively, it is crucial to be able to extract the gist of these documents. Having a Text Summarization system would thus be immensely useful in serving this need. The objective of automatic text summarization is to extract essential sentences that cover almost all the concepts of a document so that users are able to comprehend the ideas the documents tries to address by simply reading through the corresponding summary. In this paper we investigate some novel technique to develop an effective automatic Oriya text summarizer. These techniques can efficiently and effectively save users' time while summarizing a particular text.

Keywords

Information extraction, web search, Word frequency method, Positional Criteria method, Cue phrase method, Title overlap method.

1. INTRODUCTION

Information can be made more digestible in a number of ways. It can be compressed into a briefer format to enable the user to absorb the information quickly .Automatic summarization has become an important application recently due to the increased amount of information available on the Web. Summarization techniques can be very useful in improving the effectiveness of Web search. In order to generate a summary, we have to identify the most important pieces of information from the document, omitting irrelevant information and minimizing details, and assemble them into a compact coherent report. This however, is easier said than done as it involves some of the main problems of natural language processing. To produce a domain-independent system would require work in natural language understanding, semantic representation, discourse models, world knowledge, and natural language generation. Successes in domain-independent systems are few and limited to identifying key passages and sentences of the document. More successful systems have been produced for limited domain applications such as report generations for weather, financial and medical databases. In automatic text summarization there are two distinct techniques either text extraction or text abstraction. Text extraction means to extract pieces of an original text on a statistical basis or with heuristic methods and put together it to a new shorter text with the same information content. There are three steps to perform text extraction. First to understand the topic of a text, so called topic identification, secondly the interpretation of the text and finally the generation of the text. In text extraction the method is basically to give scores to each sentence depending on the importance of each sentence and when creating the summary the most significant sentences are kept. The scores can be based on high-frequent open word class words, bold or numerical text, proper nouns, citations, position in text etc. Text abstraction is to parse the original text in a linguistic way, interpret the text and find new concepts to describe the text and then generate a new shorter text with the same information content. The latter is very similar to text generation.

2. RELATED WORK

SweSum is the first automatic text summarizer for Swedish. It summarizes Swedish news text in HTML/text format on the WWW. During the summarization 5-10 key words - a mini summary is produced.Accurancy 84% at 40% summary of news with an average original length of 181 words.

Automatic text summarization is based on statistical, linguistical and heuristic methods where the summarization system calculates how often certain key words (the Swedish system has 700 000 possible Swedish entries pointing at 40 000 Swedish base key words). The key words belong to the so called open class words. The summarization system calculates the frequency of the key words in the text, which sentences they are present in, and where these sentences are in the text. It considers if the text is tagged with bold text tag, first paragraph tag or numerical values. All this information is compiled and used to summarize the original text.

Currently available major Web search engines use short summaries of document contents in displaying their results. Google creates document summaries using query-biased techniques. WebDocSum is a retrieval interface providing longer query-biased summaries to improve the search experience of Web users . The system uses surface-level extraction techniques; that is, it scores and selects sentences based on features such as title, location, relation to query and text formatting for the output summary. Most of the work so far has been focused on English and other European language.

In this paper, we describe an automatic text summarizer for Oriya language. The methods used are

1. Word frequency method

2. Positional Criteria method

3. Cue Phrase method

4. Title overlap method

3. STRUCTURAL PROCESSING

Currently, most of the documents on the Web are formatted in HTML. Usually, the documents are not in a fixed format.,They are designed as consisting of sections and subsections using a limited number of HTML formatting tags. Someof these tags include:

- Bold (<b>)
- Underlined (<u>)

• Font (<font>): together with the size attribute to specify the size of the font used

• Heading: <h1>, <h2>, <h3>, <h4>, <h5> and <h6> for different levels of headings

The structure of a document may be considered as a hierarchy, where each document has sections; each section has subsections, and so on . The sections and subsections are identified using some heuristics on HTML tags.

4. WEB PAGE SUMMARIZATION

There is no such automatic text summarizer for Oriya language.

We have developed an automatic text summarizer for Oriya language by using 4 different techniques.

The techniques used are:

Word Frequency method

Positional Criteria method

Title Overlap method

Cue Phrase method

These techniques can efficiently and effectively save users' time while summarizing a particular text. We have also applied some linguistic approach because in Oriya we found some Stop word and Cue word .We have used the morphological analyzer to find the root word. For morphological analyzer we have collected all possible suffixes of Oriya language.

4.1 Word frequency method

Luhn (1959) used Zipf 's Law of word distribution(a few words occur very often, fewer words occur somewhat often, and many words occur infrequently) to develop the following extraction criterion: if a text contains some words unusually frequently, then sentences containing these words are probably important.

The systems of Luhn (1959), Edmundson (1969), Kupiec, Pedersen, and Chen(1995), Teufel and Moens (1999), Hovy and Lin (1999), and others employ various frequency measures, and report performance of between 15 per cent and 35 per cent recall and precision (using word frequency alone). But both Kupiecet al. and Teufel and Moens show that word frequency in combination with other measures is not always better. Witbrock and Mittal (1999) compute a statistical model describing the likelihood that each individual word in the text will appear in the summary, in the context of certain features (part-of-speech tag, word length, neighbouring words,average sentence length, etc.). The generality of this method (also across languages)makes it attractive for further study.

If a text contains some words unusually frequently, then sentences containing these words are probably important.

4.2 Algorithm

The algorithm for the word frequency method is

1. Read Oriya web page

2. Take the Oriya text as input

3. Extract the title portion of the web page

4. Read the body part of the web page

5. Tokenize the whole text

6. Consider the tokens as words.

7. Compare these words with the predefined stop word list

8. If any sentence having stop words

9. Remove those stop word

8. Count the no of words present in the input text.

9. Count the no of unique words present in the input text.

10. Calculate the threshold value (total no. of words/no. of unique words).

11. Find the words having higher threshold value

12 .Consider it as the frequent word

13. Extract the whole sentence having those words

14. Order the extracted sentences as in the input text

A stop word is a type of word that appears very frequently in a text collection .It has no great significance .In the Information Retrieval (IR) community, stop words are defined as grammatical or function words. Prepositions, coordinators, determinants are the stop-words. Stop words are useless for search and retrieval purposes. We have identified 79 stop words for Oriya language.

Table1: Stopword list for Oriya language

No	Oriya word	English word	20	Oriya word	English word
1	ଗୋଟିଏ	GOTIE	21	କାହା	KAHA
2	ପରେ	PARE	22	କରି	KARI
3	ଏବେ	ABE	23	ଆମ	AAMA
4	ଉର୍ଦ୍ଧ	URDHA	24	ବୋଲି	BOLI
5	ଏହା	AEHA	25	ହେବ	HEBA
6	ଏତେ	AETE	26	ଏଠାରେ	AETHARE
7	କୌଣସି	KOUNASI	27	ଏକ	EKA
8	ଗତ	GATA	28	ଉକ୍ତ	UKATA
9	ଥିଲା	THILA	29	ଭାବେ	BHABE
10	ଅନୁସାରେ	ANUSARE	30	ପୂର୍ବରୁ	PURBAR U
11	ଅନେକ	ANEKA	31	ଆଜି	AAJI
12	ମଧ୍ୟ	MADHYA	32	ଯେଉଁ	JEUN

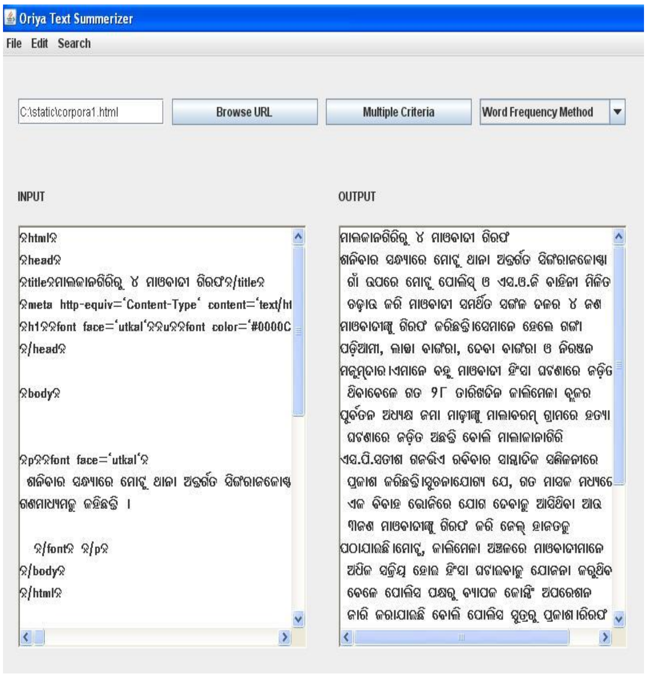
13	ପାଇଁ	PAIN	33	ସବୁ	SABU
14	ତାହାକୁ	TAHAKU	34	ପର୍ଯ୍ୟନ୍ତ	PARJYANTA
15	କେଉଁଠି	KEUNTHI	35	ସୁଦ୍ଧା	SUDDHA
16	କିନ୍ତୁ	KINTU	36	ଦ୍ୱାରା	DUARA
17	ଅନେକ	ANEKA	37	ମୋର	MORA
18	ଆଧାର	AADHARA	38	କିଏ	KIE
19	କରି	KARI	39	କନ	KANA
40	ସିଧାସଳଖ	SIDHASALAKHA	60	କିମ୍ବା	KIMBA
41	ନେଇ	NEI	61	ଏବଂ	ABANG
42	ଲାଗି	LAGI	62	ସେଥିରେ	SETHIRE
43	ଓ	O	63	ଏହି	AEHI
44	ଅନ୍ୟାନ୍ୟ	ANNAYANYA	64	ଏପରି	AEPARI
45	ଦେଇ	DEI	65	ଅବସ୍ଥା	ABASAYA
46	ଯେ	JE	66	ଏବେ	ABE
47	ଯେନୁ	JEUN	67	ତେବେ	TEBE
48	ତାହା	TAHA	68	ଟଙ୍କା	TANKA
49	ସେ	SE	69	ଏଇ	AAI
50	ଆହାର	AHARA	70	କେତେକ	KETEKA
51	ଫଳରେ	PHALARE	71	ଜଣେ	JANE
52	ଭଳି	BHALI	72	ସମସ୍ତ	SAMASTA
53	ସକାଶେ	SAKASHE	73	ତେଣୁ	TENU
54	ଅପେକ୍ଷା	APEKHYA	74	ପରିବର୍ତ୍ତେ	PARIBARTE

55	ଥିବ	THIBA	75	ଏସବୁ	ESABU
56	ଯୋଗୁ	JOGU	76	ତ୍ୱରିତ	TUARIT
57	ହେଲା	HELA	77	ଆମେ	AAME
58	ଟଙ୍କାର	TANKARA	78	ବେଳେ	BELE
59	ପାଇଁ	PAIN	79	କାରଣ	KARANA

Figure 1. Original Oriya HTML Document.

ମାଲକାନଗିରିରୁ ୪ ମାଓବାଦୀ ଗିରଫ

ମାଲକାନଗିରି , ୩୦/୩(ଡି.ଏନ.ଏ)- ଶନିବାର ସନ୍ଧ୍ୟାରେ ମଟ୍ରୁ ଥାନା ଅନ୍ତର୍ଗତ ସ୍ତଜୋଳକେ· । ଜ୍ଞୋ ଗାଁ ଉପରେ ମଟ୍ରୁ ପୋଲିସ ଓ ଏସ.ଓ.ଜି ବାହିନୀ ମିଳିତ ଚଢ଼ଉ କରି ମାଓବାଦୀ ସମର୍ଥିତ ସଂଗଠନ ଦଳର ୪ ଜଣ ମାଓବାଦୀଙ୍କୁ ଗିରଫ କରିଛନ୍ତି । ସେମାନେ ହେଲେ **ଗଙ୍ଗା ପଡ଼ିଆମି , ଲାଙ୍ଗୁ ବାଡ଼ୋ , ଦେବେ ବାଡ଼ୋ ଓ ନିରଞ୍ଜନ** ମୁଦୁଲିର । ଏମାନେ ବହୁ ମାଓବାଦୀ ହିଂସା ଘଟଣାରେ ଜଡ଼ିତ ଥିବାବେଳେ ଗତ ୨୮ ତାରିଖଦିନ କାଲିମେଳା ବ୍ଲକର ପୂର୍ବତନ ଅଧ୍ୟକ୍ଷ ଜମା ମାଡ଼ିଙ୍କୁ ମାଲାବରମ୍ ଗ୍ରାମରେ ହତ୍ୟା ଘଟଣାରେ ଜଡ଼ିତ ଅଛନ୍ତି ବୋଲି ମାଲକାନଗିରି ଏସ.ପି.ସତୀଶ ଗଜଭିଏ ରବିବାର ସାମ୍ବାଦିକ ସମ୍ମିଳନୀରେ ପ୍ରକାଶ କରିଛନ୍ତି । ସୂଚନାଯୋଗ୍ୟ ଯେ, ଗତ ମାସକ ମଧ୍ୟରେ ଏକ ବିବାହ ଭୋଜିରେ ଯୋଗ ଦେବାକୁ ଆସିଥିବା ଆଉ ୩ଜଣ ମାଓବାଦୀଙ୍କୁ ଗିରଫ କରି ଜେଲ୍ ହାଜତକୁ ପଠାଯାଇଛି । ମଟ୍ରୁ , କାଲିମେଳା ଅଞ୍ଚଳରେ ମାଓବାଦୀମାନେ ଅଧିକ ସକ୍ରିୟ ହୋଇ ହିଂସା ଘଟାଇବାକୁ ଯୋଜନା କରୁଥିବା ବେଳେ ପୋଲିସ ପକ୍ଷରୁ ବ୍ୟାପକ କମ୍ବିଙ୍ଗ ଅପରେସନ ଜାରି କରାଯାଇଛି ବୋଲି ପୋଲିସ ସୂତ୍ରରୁ ପ୍ରକାଶ । ଗିରଫ ମାଓବାଦୀମାନଙ୍କଠାରୁ ମିଳିଥିବା ତଥ୍ୟ ଅନୁସାରେ **ମାଲକାନଗିରି ଜିଲାର ବହୁ ମାଓବାଦୀ ଆନ୍ଧ୍ରପ୍ରଦେଶକୁ** ପଳାଇଛନ୍ତି । ଆନ୍ଧ୍ରପ୍ରଦେଶ ପୋଲିସ ସହ ଯୋଗାଯୋଗ ସ୍ଥାପନ କରାଯାଇ ସେମାନଙ୍କୁ ଧରିବା ଲାଗି ଏକ ପୋଲିସ ଦଳ ଆନ୍ଧ୍ରପ୍ରଦେଶକୁ ପଠାଯିବ ବୋଲି ପୋଲିସ ଅଧିକାରୀମାନେ ଗଣମାଧ୍ୟମକୁ କହିଛନ୍ତି ।



C:\static\corpora1.html

Word Frequency Method

INPUT

```
<html>
<head>
<title>ମାଲକାନଗିରିରୁ ୪ ମାଓବାଦୀ ଗିରଫ</title>
<meta http-equiv=‘Content-Type‘ content=‘text/ht
<h1><font face=‘utkal‘><u><font color=‘#0000C
</head>

<body>


<p><font face=‘utkal‘>
 ଶନିବାର ସନ୍ଧ୍ୟାରେ ମୋଟୁ ଥାନା ଅନ୍ତର୍ଗତ ସିଙ୍ଗରାଜକୋଣ୍ଡ
ଗଣମାଧ୍ୟମକୁ କହିଛନ୍ତି ।

  </font> </p>
</body>
</html>
```

OUTPUT

ମାଲକାନଗିରିରୁ ୪ ମାଓବାଦୀ ଗିରଫ
ଶନିବାର ସନ୍ଧ୍ୟାରେ ମୋଟୁ ଥାନା ଅନ୍ତର୍ଗତ ସିଙ୍ଗରାଜକୋଣ୍ଡା
ଗାଁ ଉପରେ ମୋଟୁ ପୋଲିସ୍ ଓ ଏସ.ଓ.ଜି ବାହିନୀ ମିଳିତ
ଚଢ଼ାଉ କରି ମାଓବାଦୀ ସମର୍ଥିତ ସଙ୍ଗମ ଦଳର ୪ ଜଣ
ମାଓବାଦୀଙ୍କୁ ଗିରଫ କରିଛନ୍ତି।ସେମାନେ ହେଲେ ଗଙ୍ଗା
ପଡ଼ିଆମୀ, ଲାଲା ବାଙ୍ଗରା, ଦେବା ବାଙ୍ଗରା ଓ ନିରଞ୍ଜନ
ମଜୁମ୍ଦାର।ଏମାନେ ବହୁ ମାଓବାଦୀ ହିଂସା ଘଟଣାରେ ଜଡ଼ିତ
ଥିବାବେଳେ ଗତ ୨୮ ତାରିଖଦିନ କାଲିମେଲା ବ୍ଲକର
ପୂର୍ବତନ ଅଧ୍ୟକ୍ଷ ଜମା ମାଡ଼ୀଙ୍କୁ ମାଲାବରମ୍ ଗ୍ରାମରେ ହତ୍ୟା
ଘଟଣାରେ ଜଡ଼ିତ ଅଛନ୍ତି ବୋଲି ମାଲାକାନାଗିରି
ଏସ.ପି.ସତୀଶ ଗଜଭିଏ ରବିବାର ସାମ୍ବାଦିକ ସମ୍ମିଳନୀରେ
ପ୍ରକାଶ କରିଛନ୍ତି।ସୂଚନାଯୋଗ୍ୟ ଯେ, ଗତ ମାସକ ମଧ୍ୟରେ
ଏକ ବିବାହ ଭୋଜିରେ ଯୋଗ ଦେବାକୁ ଆସିଥିବା ଆଉ
୩ଜଣ ମାଓବାଦୀଙ୍କୁ ଗିରଫ କରି ଜେଲ୍ ହାଜତକୁ
ପଠାଯାଇଛି।ମୋଟୁ, କାଲିମେଲା ଅଞ୍ଚଳରେ ମାଓବାଦୀମାନେ
ଅଧିକ ସକ୍ରିୟ ହୋଇ ହିଂସା ଘଟାଇବାକୁ ଯୋଜନା କରୁଥିବ
ବେଳେ ପୋଲିସ ପକ୍ଷରୁ ବ୍ୟାପକ କୋମ୍ବିଂ ଅପରେଶନ
ଜାରି କରାଯାଇଛି ବୋଲି ପୋଲିସ ସୂତ୍ରରୁ ପ୍ରକାଶ।ଗିରଫ

Figure 2:Output of Word Frequency method

Precision = correct / (correct + wrong)

Recall = correct / (correct + missed)

Correct = the number of sentences extracted by the system and the human;

Wrong = the number of sentences extracted by the system but not by the human

Missed = the number of sentences extracted by the human but not by the system

For this method Precision=83% Recall=83 %

4.3 Positional criteria

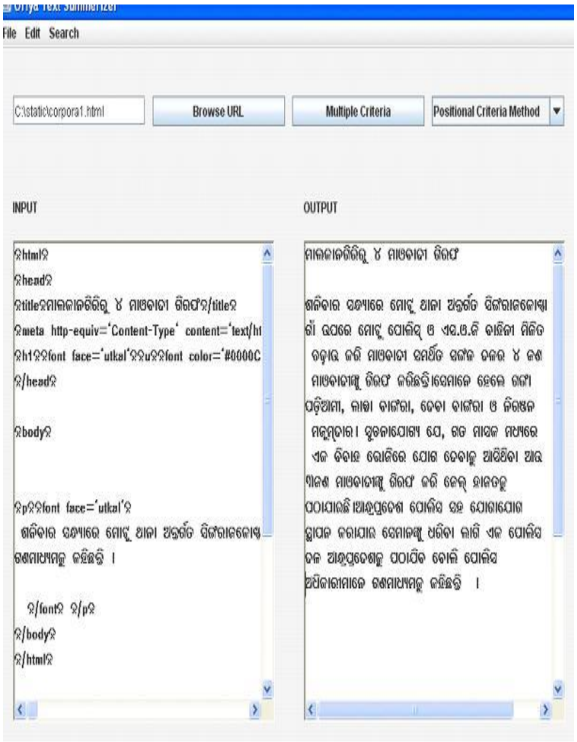
Certain locations of the text (headings, titles, first paragraphs, etc.) tend to contain important information. The simple method of taking the lead (first paragraph) as summary often outperforms other methods, especially with newspaper articles (Brandow, Mitze, and Rau1995). Some variation of the position method appears in Baxendale (1958); Edmundson (1969); Donlan (1980); Kupiec, Pedersen, and Chen (1995); Teufel and Moens (1997); Strzalkowski et al. (1999); Kupiec et al. and Teufel and Moens both list this as the single best method, scoring around 33 percent, for news, scientific, and technical articles.In order to automatically determine the best positions, and to quantify their utility,Lin and Hovy (1997) define the genre- and domain-oriented Optimum Position Policy (OPP) as a ranked list of sentence positions that on average produce the highest yields for extracts, and describe an automated procedure to create OPPs given texts and extracts.

For Oriya automatic text summarization, we are considering the first line of the first paragraph as well as the last line of the last paragraph for a given text as an important position to generate summary. We are also considering some position as important which contains the features like bold, Italic, Underline.

The algorithm for the Positional Criteria method is

1. Read Oriya web page
2. Extract the title portion of the web page
3. Read the body part of the web page
4. Tokenize the whole text
5. Extract the first line of the first paragraph
6. If any sentence having some bold letter
7. Extract the whole sentence
8. If any sentence having some italic letter
9. Extract the whole sentence
10. If any sentence having the underline
11. Extract the whole sentence
12. Extract the last line of the last paragraph
13. Order the extracted sentences as in the input text

Figure 3: Output of Positional Criteria method



C:\static\corpora1.html

Positional Criteria Method

INPUT

```
<html>
<head>
<title>ମାଲକାନଗିରିରୁ ୪ ମାଓବାଦୀ ଗିରଫ</title>
<meta http-equiv=‘Content-Type‘ content=‘text/ht
<h1><font face=‘utkal‘><u><font color=‘#0000C
</head>

<body>


<p><font face=‘utkal‘>
 ଶନିବାର ସନ୍ଧ୍ୟାରେ ମୋଟୁ ଥାନା ଅନ୍ତର୍ଗତ ସିଙ୍ଗରାଜକୋଣ୍ଡ
ଗଣମାଧ୍ୟମକୁ କହିଛନ୍ତି ।

  </font> </p>
</body>
</html>
```

OUTPUT

ମାଲକାନଗିରିରୁ ୪ ମାଓବାଦୀ ଗିରଫ

ଶନିବାର ସନ୍ଧ୍ୟାରେ ମୋଟୁ ଥାନା ଅନ୍ତର୍ଗତ ସିଙ୍ଗରାଜକୋଣ୍ଡା
ଗାଁ ଉପରେ ମୋଟୁ ପୋଲିସ୍ ଓ ଏସ.ଓ.ଜି ବାହିନୀ ମିଳିତ
ଚଢ଼ାଉ କରି ମାଓବାଦୀ ସମର୍ଥିତ ସଙ୍ଗମ ଦଳର ୪ ଜଣ
ମାଓବାଦୀଙ୍କୁ ଗିରଫ କରିଛନ୍ତି।ସେମାନେ ହେଲେ ଗଙ୍ଗା
ପଡ଼ିଆମୀ, ଲାଲା ବାଙ୍ଗରା, ଦେବା ବାଙ୍ଗରା ଓ ନିରଞ୍ଜନ
ମଜୁମ୍ଦାର। ସୂଚନାଯୋଗ୍ୟ ଯେ, ଗତ ମାସକ ମଧ୍ୟରେ
ଏକ ବିବାହ ଭୋଜିରେ ଯୋଗ ଦେବାକୁ ଆସିଥିବା ଆଉ
୩ଜଣ ମାଓବାଦୀଙ୍କୁ ଗିରଫ କରି ଜେଲ୍ ହାଜତକୁ
ପଠାଯାଇଛି।ଆନ୍ଧ୍ରପ୍ରଦେଶ ପୋଲିସ ସହ ଯୋଗାଯୋଗ
ସ୍ଥାପନ କରାଯାଇ ସେମାନଙ୍କୁ ଧରିବା ଲାଗି ଏକ ପୋଲିସ
ଦଳ ଆନ୍ଧ୍ରପ୍ରଦେଶକୁ ପଠାଯିବ ବୋଲି ପୋଲିସ
ଅଧିକାରୀମାନେ ଗଣମାଧ୍ୟମକୁ କହିଛନ୍ତି ।

For this method Precision=100%, Recall=100%

4.4 Cue phrase method

Important sentences contain cue words.

Phrases such as "in conclusion" and "note that" in some genres indicate important content. These phrases can be detected automatically by the system. The cue phrase method, which uses meta-linguistic markers (for example, “in conclusion”) to select important phrases (Edmonson 1969). The cue phrase method is based on the assumption that such phrases provide a “rhetorical” context for identifying important sentences. The source abstraction in this case is a set of cue phrases and the sentences that contain them.The algorithm for the Cue Phrase Method is

1. Read Oriya web page
2. Take the Oriya text as input

3. Extract the title portion of the web page

4. Read the body part of the web page

5. Tokenize the whole text

6. Consider the tokens as words.

7. Compare these words with the predefined cue word list

8. If any sentence having cue words

9. Extract the whole sentence

10. Other sentences will not be considered

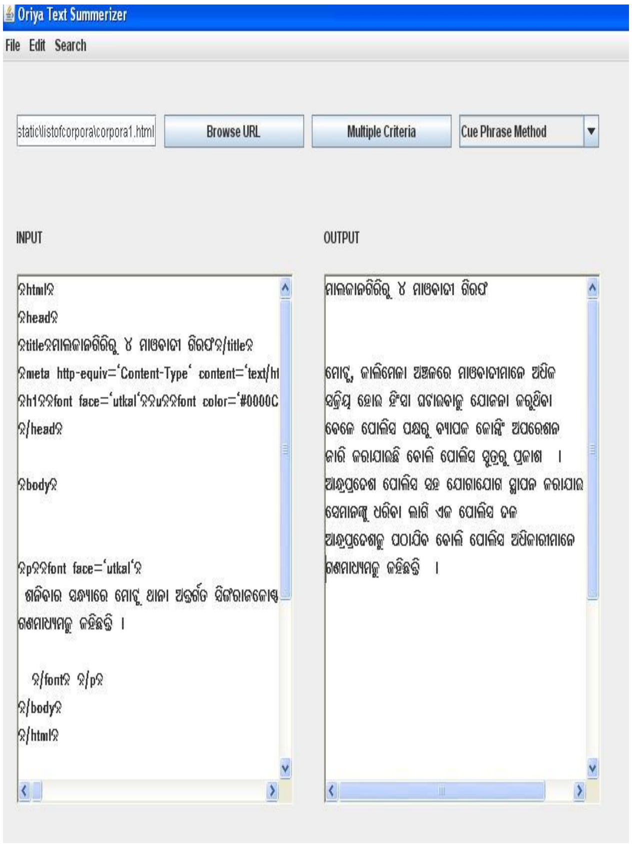
11. Order the extracted sentences as in the input text

For the cue phrase method for Oriya automatic text summarization, we have identified 37 cue words.

Table:2 Cue Word list for Oriya language

No	Oriya word	English word	No	Oriya word	English word
1	ପୁଣି	PUNI	12	ସୁତରାଂ	SUTARANG
2	ପୁନର୍ବାର	PUNARBARA	13	କାହିଁକିନା	KAHINKINA
3	ପୁନଶ୍ଚ	PUNASCHA	14	ବାରମ୍ବାର	BARAMBARA
4	କିନ୍ତୁ	KINTU	15	ପୁନଃପୁନଃ	PUNAHAPUNAHA
5	ମାତ୍ର	MATRA	16	ଯେପରିସେପରି	JEPARISEPARI
6	ତଥାପି	TATHAPI	17	ଏପରି	AEPARI
7	ପରନ୍ତୁ	PARANTU	18	ନଚେତ୍	NACHET
8	ଅଥଚ	ATHACHA	19	ନହେଲେ	NAHELE
9	ସଦା	SADA	20	କାହିଁକି	KAHINKI
10	ସର୍ବଦା	SARBADA	21	ଚମତ୍କାର	CHAMATKARA
11	ଯେବେ	JEBE	22	ଯଦି	JADI
23	କେବେ	KEBE	30	ଯେବେ	JEBE
24	ଏବେ	AEBE	31	ଯଦିଓ	JADIO
25	ତେବେ	TEBE	32	ଯଦିବା	JADIBA
26	ଯେଣୁ	JENU	33	କାଳେ	KALE
27	ତେଣୁ	TENU	34	ସ୍ୱଭାବତଃ	SWABHABATAHA
28	ଏଣୁ	AENU	35	ସାଧାରଣତଃ	SADHARANATAHA
29	ଅତଏବ	ATAEB	36 37	ଯେହେତୁ ଅଧିକନ୍ତୁ	JEHETU ADHIKENTU

Figure 4: Outpout of Cue Phrase method



static\listofcorpora\corpora1.html

INPUT

```
?html?
?head?
?title?ମାଲକାନଗିରିରୁ ୪ ମାଓବାଦୀ ଗିରଫ?/title?
?meta http-equiv='Content-Type' content='text/ht
?h1??font face='utkal'??u??font color='#0000C
?/head?

?body?


?p??font face='utkal'?
 ଶନିବାର ସନ୍ଧ୍ୟାରେ ମୋଟୁ ଥାନା ଅନ୍ତର୍ଗତ ସିଙ୍ଗାରାଜକୋଣ୍ଡ
ଗଣମାଧ୍ୟମକୁ କହିଛନ୍ତି ।

  ?/font? ?/p?
?/body?
?/html?
```

OUTPUT

ମାଲକାନଗିରିରୁ ୪ ମାଓବାଦୀ ଗିରଫ

ମୋଟୁ, କାଲିମେଳା ଅଞ୍ଚଳରେ ମାଓବାଦୀମାନେ ଅଧିକ
ସକ୍ରିୟ ହୋଇ ହିଂସା ଘଟାଇବାକୁ ଯୋଜନା କରୁଥିବା
ବେଳେ ପୋଲିସ ପକ୍ଷରୁ ବ୍ୟାପକ କୋମ୍ବିଂ ଅପରେଶନ
ଜାରି କରାଯାଇଛି ବୋଲି ପୋଲିସ ସୂତ୍ରରୁ ପ୍ରକାଶ ।
ଆନ୍ଧ୍ରପ୍ରଦେଶ ପୋଲିସ ସହ ଯୋଗାଯୋଗ ସ୍ଥାପନ କରାଯାଇ
ସେମାନଙ୍କୁ ଧରିବା ଲାଗି ଏକ ପୋଲିସ ଦଳ
ଆନ୍ଧ୍ରପ୍ରଦେଶକୁ ପଠାଯିବ ବୋଲି ପୋଲିସ ଅଧିକାରୀମାନେ
ଗଣମାଧ୍ୟମକୁ କହିଛନ୍ତି ।

For this method Precision =100%, Recall=100%

4.5 Title overlap method

A simple but useful method is to score each sentence by the number of desirable words it contains. Desirable words are, for example, those contained in the text's title or headings (Kupiec, Pedersen, and Chen 1995;Teufel and Moens 1997; Hovy and Lin 1999), or in the user's query, for a query-based summary (Buckley and Cardie 1997; Strzalkowski et al. 19999; Hovy and Lin 1999).

In title overlap method we are first extracting the title. If the title or some part of the title is present in the body section of the input text, we are also extracting those sentences after matching with the title. Matching can be done in two ways. First we are directly matching the title with the body part and the second approach is to use morphological analyzer for better accuracy.The algorithm for the Title Overlap method is

1. Read Oriya web page

2. Extract the title portion of the web page

3. Extract the body portion of the web page

4. Tokenize the whole text

5. Match the title with the body text

6. If match happens

6.1. Extract the whole sentence from the body portion

6.2 Otherwise go to the next step

7. Apply the morphological analyzer in the title

8. Find the root words from the title

9. Again use the morphological analyzer in each token of the body part

10. Find the root word

11. Take the root words of title and match with the body text

12. If the root words are matches

13. Extract those sentences

14. Compare those extracted sentences with the body part

15. Arrange them in order as in the body part

Morphological analyzer removes the suffixes and finds the root words. The suffixes may be different in the title or in the body part. But we know the root words are important.

Oriya languages are characterized by a rich system of inflections (VIBHAKTI), derivation and compound formation for which a standard Morphological Parser is needed for POS tagging. The MP system is designed according to the orthographic rules with the help of suffix table and dictionary.

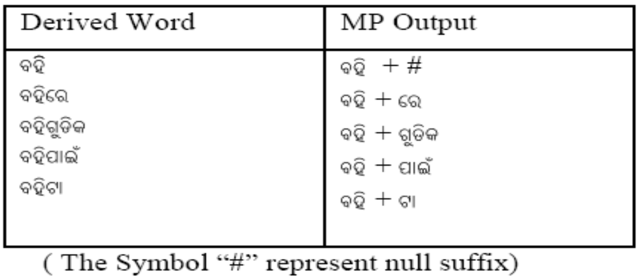
In the process of formation of words in Oriya language there exists three major classes of morpheme which are Pronoun Morphology (PM), Inflectional Morphology (IM), Derivational Morphology (DM).

Algorithm:

1. Take the Oriya text as input
2. Tokenize all the words present in the text
3. Use finite state automata(FSA)
4. Find the correct suffixes associated with the words
5. Go to the dictionary and check the availability of the word
6. Consider the word as a root word

show the output (root word + suffix)

Table 3: (Morphological Analyzer Output)



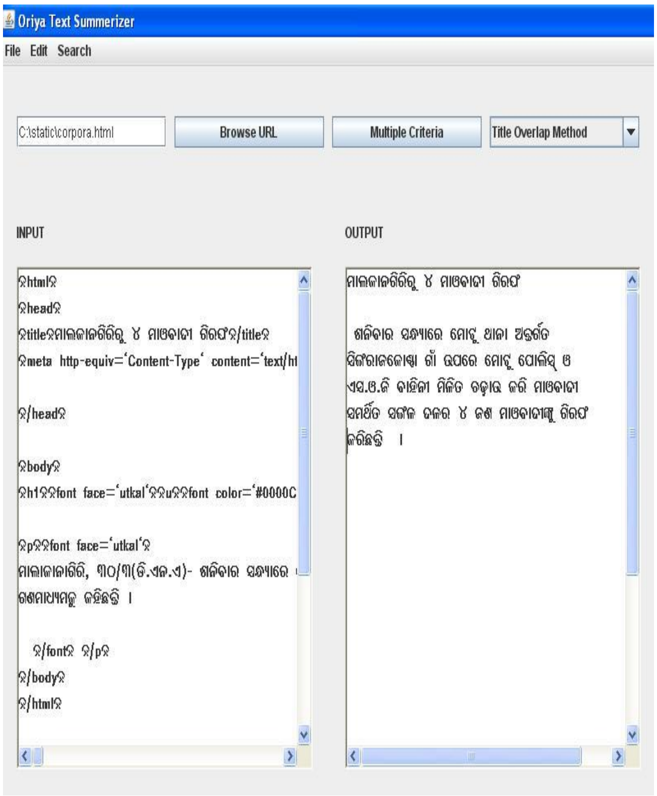
Derived Word	MP Output
ବହି ବହିରେ ବହିଗୁଡିକ ବହିପାଇଁ ବହିଟା	ବହି + # ବହି + ରେ ବହି + ଗୁଡିକ ବହି + ପାଇଁ ବହି + ଟା

(The Symbol “#” represent null suffix)

Table 4: List of suffixes for Oriya language

No	English Word	Oriya Word	No	English Word	Oriya Word
1	ucha	ଉଛ	22	ibe	ଇବେ
2	uchi	ଉଛି	23	uthiba	ଉଥିବ
3	uchu	ଉଛୁ	24	uthiba	ଉଥିବା
4	uche	ଉଛେ	25	uthibe	ଉଥିବେ
5	icha	ଇଛ	26	uthibu	ଉଥିବୁ
6	ichu	ଇଛୁ	27	uthibi	ଉଥିବି
7	iche	ଇଛେ	28	uchanti	ଉଛନ୍ତି
8	ichi	ଇଛି	29	ichanti	ଇଛନ୍ତି
9	anta	ଅନ୍ତ	30	uthanti	ଉଥନ୍ତି
10	anti	ଅନ୍ତି	31	uthantu	ଉଥନ୍ତୁ
11	antu	ଅନ୍ତୁ	32	uthanta	ଉଥନ୍ତ
12	ante	ଅନ୍ତେ	33	uthanta	ଉଥନ୍ଥ
13	anta	ଅନ୍ଥ	34	uthante	ଉଥନ୍ତେ
14	uchanti	ଉଛନ୍ତି	35	ithanta	ଇଥନ୍ତ
15	ichanti	ଇଛନ୍ତି	36	ithanta	ଇଥନ୍ଥ
16	ila	ଇଲ	37	ithanti	ଇଥନ୍ତି
17	ilu	ଇଲୁ	38	ithantu	ଇଥନ୍ତୁ

18	ili	ଇଲି	39	ithante	ଇଥାନ୍ତେ
19	ila	ଇଲା	40	e	ଇ
20	ile	ଇଲେ	41	te	ତେ
21	uthila	ଉଥିଲ	42	ta	ତା
43	uthilu	ଉଥିଲୁ	64	tae	ତାଏ
44	uthile	ଉଥିଲେ	65	taku	ତାକୁ
45	uthili	ଉଥିଲି	66	tara	ତାର
46	ithila	ଇଥିଲା	67	taru	ତାରୁ
47	ithilu	ଇଥିଲୁ	68	tare	ତାରେ
48	ithile	ଇଥିଲେ	69	ti	ତି
49	ithili	ଇଥିଲି	70	tira	ତିର
50	iba	ଇବ	71	tira	ତିରା
51	iba	ଇବା	72	tire	ତିରେ
52	ibi	ଇବି	73	tie	ତିଏ
53	ibu	ଇବୁ	74	tiku	ତିକୁ
54	tharu	ଠାରୁ	75	thare	ଠାରେ
55	thu	ଠୁ	76	gurie	ଗୁଡିଏ
56	ru	ରୁ	77	gurika	ଗୁଡିକ
57	re	ରେ	78	guriku	ଗୁଡିକୁ
58	ra	ର	79	guraka	ଗୁଡାକ
59	mana	ମାନ	80	guraku	ଗୁଡାକୁ



```
<html>
<head>
<title>ମାଲକାନଗିରିରୁ ୪ ମାଓବାଦୀ ଗିରଫ</title>
<meta http-equiv=‘Content-Type‘ content=‘text/ht

</head>

<body>
<h1><font face=‘utkal‘><u><font color=‘#0000C

<p><font face=‘utkal‘>
ମାଲକାନଗିରି, ୩୦/୩(ଡି.ଏନ.ଏ)- ଶନିବାର ସନ୍ଧ୍ୟାରେ
ଗଣମାଧ୍ୟମକୁ କହିଛନ୍ତି ।

   </font> </p>
</body>
</html>
```

```
ମାଲକାନଗିରିରୁ ୪ ମାଓବାଦୀ ଗିରଫ

 ଶନିବାର ସନ୍ଧ୍ୟାରେ ମୋଟୁ ଥାନା ଅନ୍ତର୍ଗତ
ସିଙ୍ଗାରାଜକୋଣ୍ଡା ଗାଁ ଉପରେ ମୋଟୁ ପୋଲିସ୍ ଓ
ଏସ.ଓ.ଜି ବାହିନୀ ମିଳିତ ଚଢ଼ାଉ କରି ମାଓବାଦୀ
ସମର୍ଥିତ ସଙ୍ଗଳ ଦଳର ୪ ଜଣ ମାଓବାଦୀଙ୍କୁ ଗିରଫ
କରିଛନ୍ତି  ।
```

Figure 5: Output of Title Overlap method

For this method Precision=100%, Recall=100%

5. CONCLUSION AND FUTURE WORK

In conclusion, we have presented an automatic Oriya text summarizer by using 4 novel methods. For all the methods, they show promising performance in terms of precision/recall (95%), but future work is needed in terms of the performance and robustness of the system.

In the future we would like to extend our Oriya automatic text summarizer so as to produce near-abstractive summaries. We can integrate our system with extraction based summarization system which can improve the performance of the system in future. We are also trying to introduce some other sources of knowledge including syntactic knowledge, context etc and explore other interesting application of our system.

6. REFERENCES

[1] Edmundson, H. P. 1969. ‘New methods in automatic extraction’. Journal of the ACM,16(2),264–85. Also in Mani and Maybury (1999)

[2] Liu, H. and E.H. Hovy. Automated Learning of Cue Phrases for Text Summarization.

[3] Robust Text Processing in Automated Information Retrieval-Strzalkowski-1994, ACM DBLP.

[4] C-Y Lin and E.Hovy.1997.IdentifyTopics by Position,Preceedings of the 5th Conference on applied Natural Language Processing,March.

[5] H.P.Luhn.1995.the Automatic Creation of Literature Abstracts.IBM Journal of Research and Development PP 159-165.

[6] Hovy, E. and Lin, C. Y. 1997. Automated text summarization in Summarist. In Proceedings of

ACL'97 Workshop on Intelligent, Scalable Text Summarisation, Madrid, Spain.

[7] Kupiec, J., Pedersen, J. & Chen, F. (1995). A trainable document summarizer. Proceedings of the 18th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 68-73). New York: ACM Press.

[8] Hovy, E., & Lin, C. -Y. (1999). Automated text summarization in Summarist. In I. Mani and M. Maybury (Eds.), Advances in automatic text summarization (pp. 81-94). Cambridge, MA: MIT Press.

[9] 2000 Dalianis, H. SweSum - A Text Summarizer for Swedish, Technical report TRITA-NA-P0015, IPLab-174, NADA, KTH, October 2000, html

[10] Lin, C-Y. 1997. Robust Automated Topic Identification. Ph.D. dissertation, University of Southern California.

[11] Strzalkowski, T. et al., 1998. In I. Mani and M. Maybury (eds), Advances in Automated Text Summarization. MIT Press.

[12] Lin, C-Y. 1997. Robust Automated Topic Identification. Ph.D. dissertation, University of Southern California.