Analysis of Word Frequency Distribution in Kannada Text Document

Kavya Prabhu K P[1] Divya Prabhu K P[2]
Lavanya R P[3] Vivekananda[4]
[1,2,3] 8th semester, Department of Computer Science and Engineering,
[4] Asst.Professor, Department of Computer Science and Engineering,
[1,2,3,4] Adichunchanagiri Institute of Technology,
Chikamagaluru.

***Abstract*- Summarization is the technique of reducing a text document by retaining the most important points of the original document. Summaries of any document can help to find the right information and are particularly effective when the document base is very large. The keywords that are closely associated to a document can be used to reflect the document's content.**

In this work, we propose a method to obtain the summary of any Kannada documents like press reports, fictional works etc. The input document is disassembled into its constituent words which allow us to search for well defined patterns. Later, they are categorized and processed. By determining the most frequently occurred words, the document can be summarized.

Keywords: Morphology ,keyword, word-frequency pair,summary

I. INTRODUCTION

Kannada is a Dravidian language spoken in Karnataka state of India. Kannada script is the visual form of Kannada language with a large number of structural features. It is a synthetic language of suffixing type with morphology that basically uses words which contains different morphemes to determine their meaning. Therefore, processing and summarizing of Kannada scripts is very difficult and involves several steps.

Summarization of the multiple documents is usually obtained by determining tf/idf (Term Frequency/Inverse Document Frequency) factor of every words of a document to obtain the importance of that particular word[1]. This numerical statistics helps us to obtain keywords/phrases which are closely related to the document and they reflect the contents of the document. This helps people saving their great time. But we cannot consider idf factor in summarization of a single document. So we are presenting an approach to obtain the key facts of the single document by categorizing and processing the words particularly nouns and pronouns to determine the keywords. The most frequently used keyword is used to obtain the summary of the single document.

The paper is organized as follows. Section 2 briefs about the previous work and attempts. Section 3 covers the morphological analysis. Methodology including the architecture, algorithm and implementation are detailed in section 4. The results are discussed in section 5. Section 6 gives the conclusion.

II. PREVIOUS WORK

The approach by Mari-Sanna Paukkeri et al selects words and phrases that best describe the meaning of the documents by comparing ranks of frequencies in the documents to the reference corpus. Method of You Ouyang extracted the most essential words and then expanded the identified core words as the target key phrases by word expansion approach. A novel approach to key phrase extraction proposed by them consists of two stages: identifying core words and expanding core words to key phrases.

III. KANNADA MORPHOLOGY

In linguistics, morphology deals with the study of words, their formation, relationship with other words in same language. Kannada morphology is of agglutinative type where root words are inflected with various morphemes to obtain several different words with different meanings. The words of the language are categorized into declinable words (namapada), conjugable words (kriyapada) and uninflected words (avyaya)[2]. Declinable words are inflected to depict the differences of case, number and gender as shown in table1. The conjugable words are inflected to depict the differences of gender, number, person and tense. Uninflected words are unchangeable.

Table 1: Different cases for declinable noun

Characteristic Suffix	Kannada Name	English Name	Example
u (nu/Lu/ru/vu/yu)	Prathama	Nominative	ರಾಮನು, ಸೀತಾಳು, ಮರವು
Annu/vannu/ rannu/nannu/Lannu	Dwitiya	Accusative	ರಾಮನನ್ನು,ಸೀತಾಳನ್ನು, ಮರವನ್ನು
iMda/niMda/ riMda/LiMda/diMda	Tritiya	Instrumental	ರಾಮನಿಂದ,ಸೀತಾಳಿಂದ,ಮರಗ

			ಳಿಂದ
ge/ige/kke	Chaturthi	Dative	ರಾಮನಿಗೆ, ಸೀತಾಳಿಗೆ, ಮರಕ್ಕೆ
deseyiMda	Panchami	Ablative	ರಾಮನ ದೆಸೆಯಿಂದ
a/da/ra/na/La	Shashti	Genitive	ರಾಮನ, ಸೀತಾಳ, ಮರದ
alli/valli/nalli/dalli/Lalli	Saptami	Locative	ರಾಮನಲ್ಲಿ,ಸೀತಾಳಲ್ಲಿ,ಮರದಲ್ಲಿ
Ee	Sambhodana	Vocative	ತಾಯೀ

Other than the above mentioned characteristic suffixes (cases), other words that can be attached to the declinable words while framing the sentences are ಗೋಸ್ಕರ, ಆದುದರಿಂದ, ಅವರಿಂದ, ಒಳಗೆ etc.

IV. METHODOLOGY

4.1 Architecture

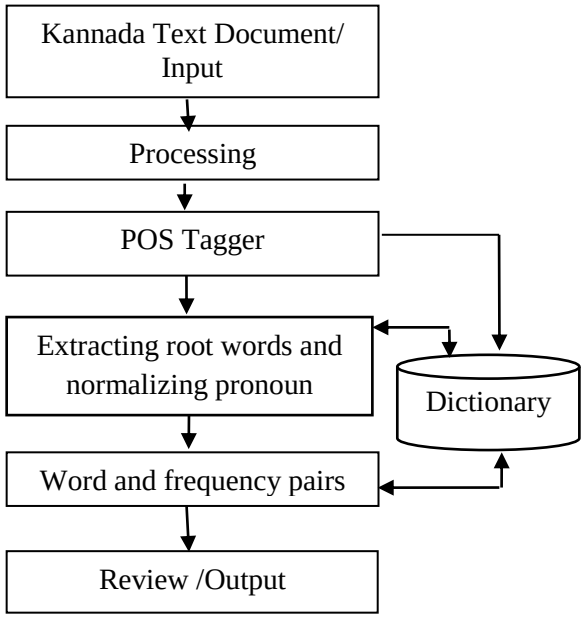

Figure 1: Architecture of document summarization

The Kannada text document is given as input from the user. This document is processed and its words are tagged with their respective parts of speech. This information is saved in the dictionary. Then the most frequently occurred keywords helps in analyzing the document. The overall architecture for summarization of document is as shown in the figure 1.

4.2 Algorithm

The algorithm used summarizes the document on the basis of the noun that has occurred maximum number of times in the given document. The following are simpler steps of the summarizing algorithm:

Algorithm: Summarizing_algorithm
Step 1: Providing Kannada text document as input.
Step 2: Tokenizing the document into sentences and words.
Step 3: Tagging of parts of speech (POS) manually.
Step 4: Extracting the root words from declinable words.
Step 5: Determining suitable noun for the pronoun and replacing it with the same.
Step 6: Determining the frequency of words (nouns and replaced pronouns).
Step 7: Analyzing the document based on keyword obtained.
Step 8: Reporting the summary of the given document

4.3 Implementation

At first, any Kannada document which contains either a single paragraph or multiple pages is provided as input. Then, the raw data of the document is subjected to processing where it is tokenized into its constituent sentences and words. The constituting words of the document are tagged with parts-of-speech by permitting the user to enter the details such as type, gender, number. This extracts the root word from the word that is tagged with noun and the words that are tagged with pronoun are replaced with the suitable noun.
For example, consider the sentence:
ರಾಮನು ವನವಾಸಕ್ಕೆ ಹೋದನು. ಅವನು ಬಹಳ ಶ್ರಮ ಪಟ್ಟನು.

Here the root word ರಾಮ is extracted from the word ರಾಮನು (suffixed with the nominative case) and ವನವಾಸ from ವನವಾಸಕ್ಕೆ (suffixed with dative case) where both the words belong to noun category.
When we encounter pronoun, careful analysis of the context of the sentences is required to substitute that word with the corresponding noun. Few cases are considered below

Case 1: If the sentence has only one subject noun and object noun of particular category. Example: ರಾಮನು ವನವಾಸಕ್ಕೆ ಹೋದನು. ಅವನು ಬಹಳ ಶ್ರಮ ಪಟ್ಟನು.

Here, the pronoun ಅವನು is replaced with the word ರಾಮ on the basis of rule that subject pronoun is replaced with former subject noun and object pronoun with former object noun.
Case 2: If the sentence has more than one subject noun and object noun with different case.
Example: ರಾಮನು ಲಕ್ಷ್ಮಣನೊಡನೆ ವನವಾಸಕ್ಕೆ ಹೋದನು. ಅವನು ಬಹಳ ಶ್ರಮ ಪಟ್ಟನು.
When we consider the above sentence, we encounter two subject nouns i.e.,ರಾಮ and ಲಕ್ಷ್ಮಣ with different

suffixes or cases such as ನು (ಉ) and ನೊಡನೆ (ಒಡನೆ) respectively . In this situation, the priority of the suffix attached is determined to know the word required to replace pronoun. As nominative case (ಉ) has higher priority over other morphological suffixes(ಒಡನೆ), the pronoun ಅವನು is replaced with the word ರಾಮ.

Case 3: If the sentence is ambiguous for the replacement of pronoun.

Example: ರಾಮನಿಗೆ ಲಕ್ಷ್ಮಣನೆಂಬ ತಮ್ಮನಿದ್ದನು. ಅವನು ತುಂಬ ತುಂಟನಾಗಿದ್ದನು. ಜೊತೆಗೆ ಅಣ್ಣನ ಮಾತು ಕೂಡ ಕೇಳುತಿರಲಿಲ್ಲ.

ರಾಮನಿಗೆ ಲಕ್ಷ್ಮಣನೆಂಬ ತಮ್ಮನಿದ್ದನು. ಅವನು ತುಂಬ ತುಂಟನಾಗಿದ್ದನು. ಆದರು ತನ್ನ ತಮ್ಮನನ್ನು ಪ್ರೀತಿಸುತಿದ್ದನು.

In the above sentences, the ambiguity arises while replacing the pronoun ಅವನು. The previous sentence alone is not sufficient to determine the appropriate Noun. Hence we start processing the next sentence which helps us to determine the noun for the particular pronoun.

After processing and replacing the pronoun, we determine the frequency of the root words of the document. The word frequency pair allows us to identify the words with maximum frequency and the most frequently occurred words are taken as keywords. The keywords are then used to analyze the document and thus overview of an document is obtained.

V. RESULTS

In this work, major emphasis is made on nouns and pronouns that appear in the document. Consider the below sentences as the input document provided.

ಪ್ರತಿಷ್ಠಾನ್ ಪುರದಲ್ಲಿ ನರ್ಮದಾ ಎಂಬ ಸತಿಯಿದ್ದಳು. ಅವಳು ಶಾಂಡಿಲೀಗೋತ್ರದಲ್ಲಿ ಹುಟ್ಟಿದುದರಿಂದ ಅವಳು ಶಾಂಡಿಲೀ ಎಂದು ಪ್ರಸಿದ್ಧಳಾದಳು. ಅವಳ ಪತಿ ಕೌಶಿಕ ಎಂಬ ಪ್ರಸಿದ್ಧ ಬ್ರಾಹ್ಮಣನು. ಅವನು ಪೂರ್ವ ಜನ್ಮದ ಪಾಪದಿಂದ ಕುಷ್ಠರೋಗಿಯಾಗಿದ್ದನು. ಸಿರಿವಂತನದುದರಿಂದ ಅವನು ಅಡ್ಡದಾರಿ ಹಿಡಿದಿದ್ದನು.

Figure 2: Input text document

Processing of nouns and pronouns gives the root words (dhatus) of the inflected words.

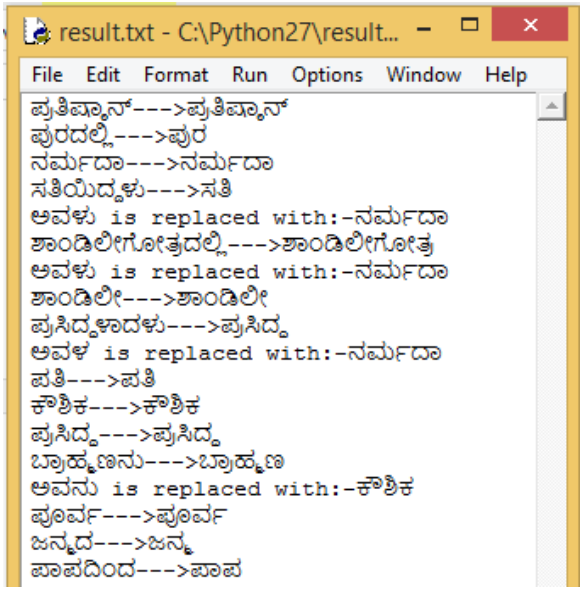

Figure 3: Processing of noun and pronoun

Output:

When the frequency of the nouns along with the replaced pronouns is determined, it is observed that the word 'Narmada' has occurred frequently and thus we can infer that the document discusses about a person Narmada.

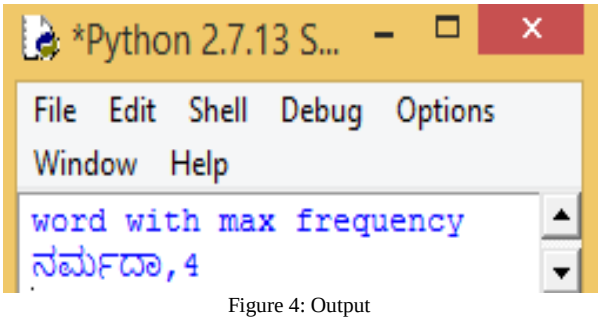

Figure 4: Output

VI. CONCLUSION

In the last two decades, there has been a revolution in the development of Indian natural language processing. Even though Kannada language is rich in literature, very less work has been carried out. So, a little effort has been made to summarize the Kannada script by concentrating only on the nouns and pronouns. But, this method cannot ensure the proper result when there occurs ambiguities because of different verb forms.

REFERENCES

[1] Jayashree R, Srikanta Murthy K, Basavraj.S.Anami: Categorized Text Document Summarization in the Kannada Language by Sentence Ranking, 12th International Conference on Intelligent Systems Design and Applications (ISDA),2012

[2] S.N. Sridhar, Modern Kannada Grammar, Manohar Publishers and Distributors, New Delhi, 2007

[3] B M Sagar, Dr Shobha G I, Dr. Ramakanath Kumar: Context Free Grammar(CFG) Analysis for Simple Kannada Sentences, Special Issue of IJCCT Vol. 1 Issue 2,3,4; 2010 for International Conference [ACCTA-2010], 3-5 August 2010