Learning Approaches toward Title Word Selection on Indic Script

P.Vijayapal Reddy
Department of Computer Science & Engineering
Raja Mahendra College of Engineering
Hyderabad,India

A.Govardhan
Department of Computer Science & Engineering
JNTUH College of Engineering,Jagityal
Karimnagar,India

Abstract—Title is a compact representation of a document which distill the important information from the document. In this paper we studied the selection words as title words by using different learning approaches namely nearest neighbor approach (NN), Naive Bayes approach with limited-vocabulary (NBL), Naive Bayes approach with full vocabulary (NBF) and by using a term weighing approach (tf-idf). We compare the performance of these approaches by using F1 metric. We compare the F1 metric results both on English Script and Indic Script ' Telugu'. We concluded the influence of linguistic complexity in the process of Title word selection.

Keywords-Title; F1 measure; NN approach; NBL approach; NBF approach; tf-idf approach;

I. INTRODUCTION

A title is a compact representation of a document which contains document's main theme, so that readers can quickly identify the information that is of interest to them. Title of a document is distinctively different from abstract of a document. Titles represents most important theme of the input text while abstracts use relatively more words and reflect many important points of the input text [XII] . The process of automatic title generation (ATG)needs to have the knowledge about the content of a document and the knowledge to create a title in a human readable sequence [I] that actually reflects the content in only a few words. This specific nature distinguishes automatic title generation from text summarization [VII] and information retrieval [VIII].

Automatic title generation, can be used for different applications such as summarizing emails and web pages etc.. for mobile phones and PDAs,to generate titles for retrieved documents by most commercial search engines. Also ATG can be used to create titles for speech recognition transcripts, machine-translated documents. ATG can also be used create titles in one language where as the documents are written in another language which is known as cross-lingual title generation,so that it can be quite useful to cross-lingual information retrieval task.

Automatic title generation approaches can be broadly divided into two categories such as text summarization based approaches and Statistical learning approaches [XII]. In text summarization based approaches title can be treated as a summary with very short length and can apply these approaches directly on to the document to get the title. We can use directly the existing methods in the field of text summarization for title generation. The quality of generated title becomes very poor when the compression on the summary of a document falls below a specific threshold [IV].

Statistical approaches are based on learning the relation between the title and the document from training corpus, and based on this knowledge title can be created for test document. These approaches can be easily applied to different domains and to different languages. Statistical based approaches can be used for cross-lingual title generation i.e. document is in one language where as its title can be available in another language as in [II]. The performance of statistical approaches heavily depends on training corpus size. Statistical based approaches requires to find the relation between title and document words leads to utilization more computational resources. In this we have evaluated the different Statistical approaches for title word selection and compared the results between English and Telugu corpus.

The outline of this paper is as follows: Section 1 gave an introduction to the title generation problem. In Section 2 we presented the detailed procedure of each statistical method. The details of experimental design and about evaluation measure are explained in section 3. The comparative results and discussion about different approaches are presented in section 4. The conclusions and future scope of the work is explained in section 5.

II. DESCRIPTION ABOUT STATISTICAL APPROACHES

A. NAIVE BAYES APPROCH WITH LIMITED-VOCABULARY (NBL) APPROACH

Naïve Bayes approach for title word selection with limited-vocabulary (NBL) follows as in [I]. In this

Learning step: Words in the given document are selected as title words for that document based on training corpus document titles. i.e. Document is not going to have new words which are not appear in the document as title words. Probability of a word 'dw' to act as title word 'tw' can be calculated as follows:

$$P\left(t_w / d_w\right) = \frac{P\left(t_w \in title \wedge d_w \in document\right)}{P\left(d_w \in document\right)} \quad if \quad t_w = d_w$$

$$P\left(t_w / d_w\right) = 0 \quad if \quad t_w \neq d_w$$

The scores of the words are calculated by multiplying the probability of a word with its frequency.

$$S\left(t_w / d_w\right) = TF\left(t_w / d_w\right).P\left(t_w / d_w\right)$$

Selection step: Once the probability of all the words which are in the document are calculated then select first six words which are having highest scores as the average length of the training document title words is six.

B. NAIVE BAYES APPROCH WITH FULL-VOCABULARY (NBF) APPROACH

Learning step: The probability of the words in the document can be calculated by taking all possible combinations with all the words which acts as title words in the training documents. Let probability of word 'dw' to act as a title word 'tw' can be calculated as follows:

$$P\left(t_w / d_w\right) = \frac{P\left(t_w \in title \wedge d_w \in document\right)}{P\left(d_w \in document\right)}$$

The scores are calculated by multiplying the probability with its word frequency.

$$S\left(t_w / d_w\right) = TF\left(t_w / d_w\right).P\left(t_w / d_w\right)$$

Selection Step: Select first six words which are having highest scores.

C. TERM WEIGHING (TF-IDF) APPROACH

Learning step: Calculate frequency of the word within in the document (TF) and then normalize the length of the document by dividing the term frequency with the number of words (NTF) within the document as follows:

$$NTF\left(d_w\right) = \frac{TF\left(d_w\right)}{totalnumberofwords}$$

Calculate inverse document frequency (IDF) to identify word importance when compared with other document in the corpus as follows:

$$IDF\left(d_w\right) = \frac{totalnumberofdocuments}{numberofdocuments \quad containing \quad 'dw'}$$

then multiply these two values to get the importance of a word 'dw' within the document and on whole corpus as follows:

$$TF - IDF\left(d_w\right) = NTF\left(d_w\right).IDF\left(d_w\right)$$

Selection step: consider first six words which are having highest scores (TF-IDF) as the average number of title words in the training corpus is six.

D. NEAREST NEIGHBOR (NN) APPROACH

Learning step: In Nearest Neighbor approach, identify one of the documents in the training which are more similar to the given document by assigning weights to the words in the both test document and training documents as in [XII].

The number of words in each document vary, so it is required to normalize the document length. Normalization can be performed by dividing the term frequency of each word with root of sum of squares of term frequencies of each individual word. Then the normalized frequency (NTF) is calculated as follows :

$$NTF\left(d_w\right) = \frac{TF\left(d_w\right)}{\sqrt{\left(\sum\left(squaresofTFofallwords\right)\right)}}$$

Then calculate the inverse document frequency (IDF) for each word which are in both test document and training documents as follows:

$$IDF\left(d_w\right) = \frac{totalnumberofdocuments}{numberofdocuments \quad containing \quad 'dw'}$$

The weight of the word is calculated as product of normalized term frequency into inverse document frequency as follows:

$$WW\left(d_w\right) = NTF\left(d_w\right).IDF\left(d_w\right)$$

Then the similarity is performed as the dot product between two document vectors.

Selection step: Once the dot product performed between test document and training documents , select the training document title as the title of test document whose dot product results high.

III. TITLE WORD SELECTION EXPERIMENT

In this Section we presented the description about corpus collection and how test data and training data are separated from the collected corpus and at subsection A. In the section B explanation about F1 metric presented.

A. EXPERIMENT DESIGN

The experimental dataset was gathered from Famous Telugu News Papers ' Eenadu ' and ' Sakshi ' from the web during the year 2010 – 2011 in the unicode format. There are a total of 3000 documents and corresponding titles in the corpus. The training dataset is formed by picking three document-title pairs from every four pairs in the original corpus. Thus, the size of training corpus was 2250 documents with corresponding titles. The remaining 750 documents are used for testing. By separating training and test dataset in this way, we ensure a overlap in the topic content between the training set and the test set, which gives the statistical learning algorithms a chance to play a significant role in the title generation process.

B. EVALUATION METRIC

In this paper, we measure the quality of selected title words using the BMW approach by comparing with the human assigned title words. More specifically, we In this paper, we which have been broadly used in the field of Information Retrieval and which has been proved a good evaluation metric[13] to measure the quality of selected title words.
The F1 measure can be calculated by using precision and recall as in following equation.

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

where, precision is the number of common words in machine generated title $T_{machine}$ and human-generated title T_{human} divided by length of machine-generated title $T_{machine}$ as in following equation:

$$precision = \frac{T_{machine} \wedge T_{human}}{T_{machine}}$$

recall is defined as the number of common words in machine-generated title $T_{machine}$ and human-generated title T_{human} divided by the human-generated title T_{human} as in following equation:

$$recall = \frac{T_{machine} \wedge T_{human}}{T_{human}}$$

T_{human} represents the human generated title, where as T_{human} represents the machine generated title. Precision shows, in the title generated by computer, the percentage of words being "correct". Meanwhile recall gives the percentage of "correct" words that computer has selected, among the title assigned by human subjects. F1 measure balances both precision and recall measures. The First six title words having Highest scores were selected , as the average number of title words for training documents in the training corpus is six.

IV. RESULTS AND DISCUSSION

When compare the F1 measures of different approaches applied on both English and Telugu documents as shown in figure 4.1.

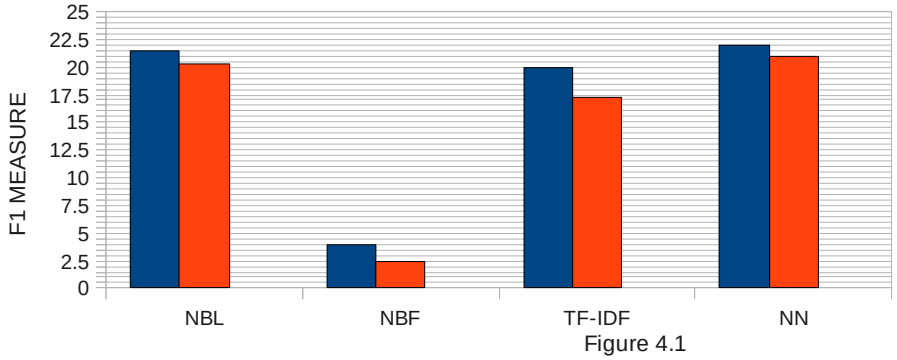

Figure 4.1

In the above figure Blue color bar shows the English corpus F1 measure percentage, where as red color bar shows the Telugu corpus F1 measure percentage. *w*hen comparing the F1 measures of NBL approach on English and Telugu datasets are 21.5 % and 20.2 % respectively. In this approach **,** in NBF approach F1 measures on English and Telugu corpus are 4.0 % and 2.38 % receptively, In TF-IDF approach the F1 measures are 19.9% and 17.2% on English and Telugu corpus and in NN approach 21.9 % and 21.0 % are on English and Telugu datasets.

NBL approach performed well when compared with with NBF and TF-IDF approach because it limits selection of title words from the original document although this restriction will not allow NBL approach to apply for cross language title generation. NBF approach which is generalization of NBL approach which allows to select any title word in the training corpus as a title word for the original document but the evaluation measure degraded drastically because of assigning equal importance to all the words though it is content word and trivial word. TF-IDF measure gives better performance when compared with NBF and slightly low performance compared with NBL. In this approach more weight has assigned to content words when compared with common words. When we observe the NN approach it is performing very well when comparing with all other approaches mainly because of strong overlap between training dataset and Test dataset among the contents of the documents. NN approach simply assigns title of one of the training documents to test document.

V. CONCLUSIONS AND FUTURE SCOPE

Telugu language has the third largest number of native speakers in India is 13th in the Ethnologue list of most-spoken languages world wide. The number of text documents available on the web and in the automation process of Government, the collection text documents increasing enormously day by day. Hence analysis of text documents written in Telugu language for different applications is mandatory. In this paper we analyzed the different statistical approaches on Telugu corpus and compared the results with English corpus. We conclude that Telugu has more complex morphological variations when compared with English so that content words belongs to same root word are distributed as more words when compared with English leads to less F1 measure.

As a future work,Telugu is complex morphological language, and it has more morphological variations when compared with the language 'English' , to increase the title word selection efficiency i.e. To represent the content of the document in the form of a title words more effectively , a detailed study is in the angle of morphological variations required. The problem of Title generation is to be addressed. Different machine learning approaches to be addressed to get more appropriate title words for the given document. The deficiency of BMW approach to be addressed. The impact of common words on Headline generation to be addressed.

REFERENCES

[1] Ultra-Summarization: A Statistical Approach to Generating Highly Condensed Non-Extractive Summaries. Michael Witbrock and Vibhu Mittal, Just Research. In Proceedings of SIGIR 99, Berkeley, CA, August 199

[2] Rong Jin and Alexander G. Hauptmann. Title generation using a training corpus.In CICLing '01: Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing, pages 208–215, London, UK,2001. Springer-Verlag

[3] E. Firmin & M.J. Chrzanowski (1999). An evaluation of automatic text summarization. In I. Mani and M. Maybury, editors. Advances in Automatic Text Summarization. MIT Press, Cambridge, Massachusetts, 1999

[4] C. H. Leung & W.K. Kan (1997). A statistical learning approach to automatic indexing of controlled index terms. Journal of the American Society for Information Science, 48 (1), 55-66, 1997.

[5] P.D. Turney (2000). Learning algorithms for keyphrase extraction. Information Retrieval, 2(4): 303-336, 2000

[6] Mani & M. Maybury (1999). Advances in Automated Text Summarization.Cambridge, MA: MIT Press, 1999

[7] K. S. Jones & P. Willett (1997). Reading in Information Retrieval. Morgan Kaufmann Publishers, 1997

[8] MUC-6 (1995), Proceeding of The Sixth Message Understanding Conference, 1995

[9] Padmaja Rani B., Vishnu Vardhan B., Kanaka Durga A., Govardhan A., Pratap Reddy L., and Vinaya Babu A. Telugu Document Classification using Baye's Probabilistic Model Technology spectrum, Journal of JNTU, vol.2 No.1, 2008, pp.26- 30

[10] M. Banko, V. Mittal, and M. Witbrock. Headline generation based on statistical translation. In the Proceedings of Association for Computational Linguistics, 2000.

[11] V. Rjiesbergen (1979). Information Retrieval. Chapter 7. Butterworths, London, 1979.

[12] Statistical Approaches toward title generation by Rong Jin , 2003, Ph.D Thesis