Analysis of BMW Model for Title Word Selection on Indic Script

P.Vijayapal Reddy
Professor
Department of CSE
Raja Mahendra Engg.College
Hyderabad,India

Dr.B. Vishnu vardhan
Professor
Department of CSE
JNTU College of
Engg.Jagityal,India

Dr.A.Govardhan
Professor
Department of CSE
JNTU College of
Engg.Jagityal,India

ABSTRACT

A title is a short summary that represents document's main theme. Title can help the reader to have the main idea without reading the entire document. To generate a title for a document, we have to select appropriate words as title words and put them in sequence. The process of generating title for a given document by using machine, can be done by using summarization approaches or by using Statistical approaches or by combing both. For a given document, selecting appropriate words for generating a title by using any available approach mainly depends on the characteristics of the language. In this paper ,we have examined the influence of the language characteristics in the process of title word selection by using the Naïve Bayes probabilistic approach (called BMW Model) on the documents which are available in the language ' Telugu '. And also we have investigated the influence of word weight for the selection of title words in BMW Model. By using F1 metric, we have evaluated the title word selection process.

General Terms

Natural Language Processing ,Machine Intelligence

Keywords

BMW Model, Indic Script, Title Word Selection, F1 measure, Statistical Approach

1.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 [13] . 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 [1] that actually reflects the content in only a few words. This specific nature distinguishes automatic title generation from text summarization [7] and information retrieval [8].

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 [13]. 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 [4].

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 fo ATG with small modifications [13]. 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. 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.

Telugu is a Dravidian language. It is the official language of Andhra Pradesh, one of the largest states of India. Telugu language has the third largest number of native speakers in india (74 millions) and is 13th in the Ethnologue list of most-spoken languages world wide. Telugu is a complex morphological variant language in which it contains more number of morphological variations for the words when compared with the language English [10]. One of the major issues related to Indic scripts is linked with the heuristic grammar rules on individual words in the representative form of pluralities, present and past tenses etc. The resultant is found to be multiplicity in vocabulary.

The basic philosophy of language representation reflecting the phonetic sequences is the basis for the complexity involved in Telugu. More amount of complexity is involved when a large set of grammar rules are applied to combine two or three words formulating into a single word. Supervised learning strategies are yet to be explored under the pretext of linguistic knowledge exploration. The individual syllable, which is the representative structure of the canonical structure ((C)C)CV [10] , posses many to one correspondence in the form of code sequences.

From a research point of view based on the language characteristics , Title generation offers plenty of challenges in Natural language processing and Natural language synthesis

2.BMW MODEL FOR TITLE GENERATION

The first statistical framework for Automatic title generation was proposed by Banko, Mittal and Witbrock[11]. In this paper, we refer it as the 'BMW model'. In BMW model the title generation task is divided into the two phases i.e., in the first phase, selecting appropriate words which are suitable as title words known as 'content selection phase'.In the second phase, organizing the selected title words into a sequence. This phase is known as surface realization phase (Title word ordering phase) [1]. According to BMW model the whole process of title generation can be represented in the form of equation as follows:

$$P\left(w_{1,} w_{2,. \cdot \cdot}, w_m / D\right) = \prod_{i=1}^{m} P\left(w_i \in T / w_i \in D\right) . P\left(len\left(T\right)=m\right) . \prod_{i=2}^{m} P\left(w_i / w_{1,. \cdot \cdot}, w_i - 1\right) \quad 2.1$$

where 'T' represents a title for the document D. T can be written as set of words $\left(w_{1,} w_{2,. \cdot \cdot} w_m\right)$ where ' m ' is the length of the title. The term $P\left(w_{1,} w_{2,. \cdot \cdot}, w_m / D\right)$ represents the probability selecting word sequence $\left(w_{1,} w_{2,. \cdot \cdot} w_m\right)$ as title T. $P\left(w_i \in T / w_i \in D\right)$ represents the probability of selecting the word w_i as title word from the document D.P(len(T)=m) is the probability of having the title length m. $P\left(w_i / w_{1,. \cdot \cdot}, w_i - 1\right)$ represents the probability of having the w_i after w_{i-1} in the title T. The goal of the model is to find for which set of words the sequence, $P\left(w_{1,} w_{2,. \cdot \cdot}, w_m / D\right)$ is maximum.

In order to estimate the appropriateness of a word w_i as a title word for the document D i.e finding $P\left(w_i \in T / w_i \in D\right)$ using Naïve Bayes approach as in equation 2.2

$$P\left(w_i \in T / w_i \in D\right) = \frac{P\left(w_i \in T \wedge w_i \in D\right)}{P\left(w_i \in D\right)} \quad 2.2$$

According to the equation 2.2, probability of selecting a word as a title word can be calculated by simply count how many documents have word w_i in the document D and in its corresponding title T, and divide it by the number of documents containing word w_i and use this ratio as the approximation for $P\left(w_i \in T / w_i \in D\right)$. To calculate $P\left(w_i / w_{1,. \cdot \cdot}, w_i - 1\right)$, bigram statistical language model is used to order the chosen title words into the sequence [3].

From the above description it is observed that, the performance of title word selection method influenced by the constraint $P\left(w_i \in T / w_i \in D\right)$ which constrains the choice of title words and does not allow words outside of the document to be used as words in the title. This limitation prevents the work from being applied to cross lingual title generation in which titles and documents are written in different languages.

3. BMW MODEL ON INDIC SCRIPT

In this paper we have evaluated the BMW model on Indic script ' Telugu ' to understand the influence of characteristics of the language in the process selecting Title words as a part of Title generation process. And also we have studied the influence word weight in process of selecting the title words for the document ' D ' .

3.1 Selection of title words

In the BMW model , process of title generation T for the document D consists of two phases. Namely title word selection phase and title word ordering phase. Title word selection can be done by using Naïve Bayes probabilistic approach as already discussed in 2.2.According to equation 2.2, we can simply count how many documents have word w in their titles and in its body, and divide it by the number of documents containing word w in their bodies and use the ratio as the approximation for $P\left(w_i \in T / w_i \in D\right)$.

3.2 Influence of word frequency for title word selection

We have studied the influence of word frequency as word weight ' wf ' as a part word weighting scheme for title word selection problem[3]. The constraint in BMW that which automatically eliminates the common words from the word selection process, such that only the content words are going to participate in the word selection process. Hence it is required to see the influence of the word weight in the word selection.

CASE 1: Calculate the score for the word w_i without considering the word weight. i.e. Score is calculated as in equation 2.2. where , $P(w \in T \wedge w \in D)$ represents count of word ' w ' occurs both at training document body and at its title in the training corpus. $P\ (w \in D)$ represents count of word ' w ' occurs at training document body in the training corpus. Then the words with highest scores will be selected as title words for the document D.

CASE 2: Calculate the score for the word w_i , by considering the frequency of the word within the document. i.e. Score can be calculated as in equation 2.4

$$P\left(w_i \in T / w_i \in D\right) = WW\left(w_i, D\right) . \frac{P\left(w_i \in T \wedge w_i \in D\right)}{P\left(w_i \in D\right)} \quad 3.1$$

where , $P\ (w \in T \wedge w \in D)$ represents count of word ' w ' occurs both at training document body and at its title in the training corpus. $P\ (w \in D)$ represents count of word ' w ' occurs at training document body in the training corpus. WW(wi , D) is the number of occurrence of document word w_i in the document D. Then the words with highest score will be selected as title words for the document D.

4. EXPERIMENTAL 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.

5. RESULTS

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 use the F1 metric 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 equation 5.1

$$F1 = \frac{2 * precision * recall}{precision + recall} \quad (5.1)$$

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 equation (5.2)

$$precision = \frac{T_{machine} \wedge T_{human}}{T_{machine}} \quad (5.2)$$

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 equation (5.3)

$$recall = \frac{T_{machine} \wedge T_{human}}{T_{human}} \quad (5.3)$$

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.

5.1 F1 measure for title word selection

We have calculated the F1 measure on our training corpus for set of test documents. by using equation 2.2.Then selected the first six high score words a title words

According to [13] F1 measure for title word selection on English corpus is 21.5 percent, where as based on our experiment F1 measure for title word selection on Telugu corpus is 20. 2 percent. The values are shown in figure 5.1, where ETWS represents English title word selection and TTWS represents Telugu title word selection.

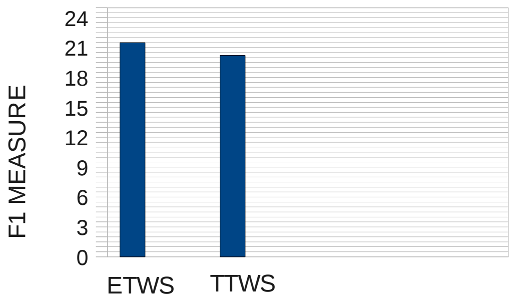

Figure 5.1

5.2 F1 measure to find influence of word weight

As in case 1 described in 3.2 without considering the word weight **the** F1 measure for title word selection according BMW model is 17. 3 percent, where as in case 2 described in 3.2 with considering word weight (ww) for title word selection F1 measure is 20.2 percent. The values are shown in figure 5.2.

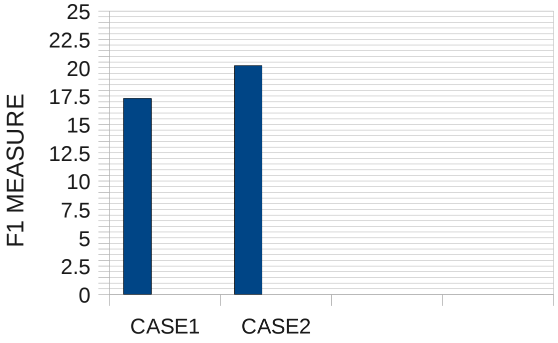

Figure 5.2

The example title words generated by without considering word weight, with considering word weight and reference title words for sample of five documents are presented in table 5.1,table 5.2 and table 5.3 respectively.

Table 5.1 Example titles words generated by the BMW method without considering the word weight

ID	Title Words Without Word Weight
1	ప్రత్యామ్నాయం లోక్ ప్రజాపథం లక్ష్మణ్ అవినీతి సమావేశం
2	కలిసికట్టుగా కలలు ఎస్సీ ఉద్యోగుల రాజ్యసభ చేసిన
3	ధర్నా వెనక్కు తెలంగాణ పార్లమెంటులో ప్రత్యేక డిమాండ్
4	నిద్ర పల్లెనిద్ర నెలా తెలంగాణ నియోజకవర్గ కు
5	వామపక్షాల భారత్ సమావేశం సీపీఎం కు విదేశీ

Table 5.2 Example titles words generated by the BMW method with considering the word weight

ID	Title Words With Word Weight
1	రచ్చబండ ప్రత్యామ్నాయం లోక్ ప్రజాపథం లక్ష్మణ్ అవినీతి
2	వైఎస్ కలిసికట్టుగా కలలు ఎస్సీ రెడ్డి ఉద్యోగుల
3	తెలంగాణ ధర్నా వెనక్కు పార్లమెంటులో ప్రత్యేక డిమాండ్
4	నిద్ర పల్లెనిద్ర తెలంగాణ టిఆర్ఎస్ నియోజకవర్గ కు
5	వామపక్షాల భారత్ సమావేశం సమావేశం సీపీఎం కు

Table 5.3 Examples of Reference Tiles

ID	Reference Title Words
1	రచ్చబండ ప్రత్యామ్నాయం కాదు
2	వైఎస్ కలలు నిజం చేద్దాం
3	ఢిల్లీలో ధర్నా చేసిన తెలంగాణ మిత్రులు
4	నేటినుంచి టిఆర్ఎస్ పల్లెనిద్ర
5	త్వరలో యూపీఏ వామపక్షాల సమావేశం

6. CONCLUSIONS & 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 investigated the problem of title word selection in the process of title generation for a given text document by using BMW approach. And also we try to explore the impact of word weight on Title word selection by using BMW approach. when compared, F1 measure on Telugu corpus is 1.3 percent less than the F1 measure on English corpus due to 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. Similarly the study on impact word weight on title word selection shows that considering the word frequency of content words in selecting title words gives more appropriate title words than without considering word frequency of content words.

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.

7. 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] Term-weighting appraoches in automatic text retrieval ,Salton and Buckley Information Processing & Management Vol. 24, No. 5, pp. 513-523,printed in Great Britain. 988

[4] 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

[5] 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.

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

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

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

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

[10] 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

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

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

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