2014 1st International Congress on Computer, Electronics, Electrical, and Communication Engineering (ICCEECE2014)
IPCSIT vol. 59 (2014) © (2014) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2014.V59.13

Automatic Headline Generation Using Context Free Grammars

Krishaprasad P, Vinayachandran KK

University Of Calicut, India

Abstract. This paper presents a novel way for generating the headline by exploiting the benefits of context free grammar. The system is based on summarizing the given document in order to get condensed form of text and then the content words are identified. Input for sentence generation, is obtained by separating named entities, nouns, verbs etc. from the content words. The system effectively generates the summary of the text document based word-frequency based scoring technique. For generating title, the paper present a context free grammar which produce suitable sentence from given content words. The experiments showed that the Title generated is efficient and the suggested titles are really helpful in extracting the important document.

Keywords: NLP, CFG, NER, Content Word Extraction

1. Introduction

This Natural Language Processing (NLP) has received a great deal of attention in recent research because of is wide applicability. Research on automatic text summarization provides basis for research on headline generation. The rapid growth of the Internet has resulted in enormous amounts of information that has become increasingly more difficult to access efficiently. The ability to summarize information automatically and present results to the end user in a compressed, yet complete form would help to solve this problem. A headline of a text, specially an article, is a succinct representation of relevant points of the input text. It differs from the task of producing abstracts, in the size of the generated text and focuses on the compressing [1] the output. Headlines are terse while abstracts are expressed using relatively more words. While headlines focus on pointing out the most relevant theme expressed in the input text, abstracts summarize the important points. This makes both headline generation and summarization extensively valuable.

Early, the problem of generating headlines for documents and text summarization uses purely statistical extraction based. Most of the summarization work done till date is based on extraction of sentences from the original document. The sentence extraction techniques compute score for each sentence based on features such as position [2] of sentence in the document, word or phrase frequency, key phrases (terms which indicate the importance of the sentence towards summary).

There were some attempts to use machine learning (to identify important features), use natural language processing (to identify key passages or to use relationship between words rather than bag of words).vector representation model is also used in text summarization techniques. Headlines are commonly associated with news articles but have wide range of applications. An application area of headline generation involves generating table of contents for a document to providing support for interactive query refinement in search engines. Headlines extracted from search result web pages can be used to augment user search query. The resultant query can be used to further re-rank and improve upon the search results. This approach of augmenting a user query with keywords extracted from text being increasingly used in Contextual text search and Information Retrieval. Automatic Headline generation tries to automate the process of providing more relevant or reflective insight into the input text rather than producing catchy lines. Automating in this context has to involve has to involve some form of learning rather than algorithmic approach given the potentially

infinite stretch of natural language text. Many machine learning techniques have been explored involving varying degrees of use of natural language understanding techniques.

Context Free Grammars (CFG) is relatively recent techniques used in natural language processing. In this work we are trying to make natural headline using context free grammar. The advantage of using CFGs is that it does not require any training data. We also present a summarization technique based on sentence scoring method as part of content word extraction.

2. CFG Model For Headline Generation

The model composed of two parts. The first part is comprised of summarization [2] part and second part is headline synthesis. The algorithm for headline synthesis purely depends on input text document only. Natural language sentence generation is the heart of the algorithm

2.1. Summarization

The summarization system has both text analysis component and summary generation component. The text analysis component used to identify the features associated with each sentence. Before the extraction process text normalization is performed (The text normalization involves splitting the text into sentences). After text normalization the normalized text is passed through a feature extraction module. Feature extraction include extracting features associated with the sentences and the features associated with words such as named entities, word frequency, characters per word etc. Later in-order to summarize the system calculates the score for each sentence based on the features that we already identified in the previous step. Sentence refinement is done on the sentences with high score, and the resulting sentences are selected for the summary in the same order as they were found in the input text document. Various steps in summarization can be summarized as follows:

2.1.1. Sentence marking:

This module divides the document into sentences. It is appearing that using end-of-sentence punctuation marks, such as periods, question marks, and exclamation points, is sufficient for marking the sentence boundaries. It should be noted exclamation point and question mark are somewhat less ambiguous. Whereas periods can be appeared in non standard words like web URL (Uniform Resource Locater) s, emails etc.

2.1.2. Feature extraction:

The system extracts both the sentence level and word level features. We are actually interested only in word level features because, we need not require high quality summary for title generation process. Our aim to supply a brief summary input to headline synthesis part. For picking out best sentence from given document we follow the sentence scoring technique based on word frequency[5] and average number of characters per sentence.

2.1.3. Summary Generation:

Summary generation include tasks such as calculating the score for each sentence, selecting the sentences with high score, and refinement of the selected collection of sentences.

2.1.4. Sentence Ranking:

Summarization system follow simple but efficient sentence ranking technique based on word frequency of particular word in the sentence and average number of characters in each word. Mathematical model of sentence ranking is discussed later in this paper.

2.1.5. Sentence Selection:

After the sentences are scored, we need to select the sentences that make good summary. One strategy is to pick the top N sentence towards the summary, but this creates the problem of coherence. The selection of sentence is dependent upon the type of the summary requested. The sentences are selected based on the percentage of output text required with respect to the input document.

2.2. Headline synthesis

The Headline Synthesis involves generating a suitable headline for given input text file based on the content words extracted from the document. It comprised of three components:

2.2.1. Extracting content words:

Content words are the word which represents over all text. Identification of content words have special importance, because the quality of the title generated will depends on the exact identification of the content words.

2.2.2. **Identification of Elements for title generation**:

Analyzing the headlines, it is noted that the headline is formed by named entities or/and frequent nouns or verbs. Out of number of selected content words we are actually interested only in nouns, verbs, and named entities.

2.2.3. **Generation of Headline:**

The heart of this paper lays on fact that, context free Grammars are used for generating title from identified elements. The effectiveness of the title depends on how well natural language sentence is generated from the identified headline elements and the following refinement. It is worth to note that the model entirely depends only on given input text file. The key advantage of this headline synthesis model is that system doesn't need any separate learning process.

3. Mathematical Modeling

3.1. HS Algorithm

Through this we implementing a new algorithm named HS (Headline Synthesis) algorithm. We have already mentioned that the context free grammar based algorithm is basis for headline synthesis. The algorithm can implement as - summarizing given text from which generating the Headline. The abstract view of the algorithm is given bellow:

3.1.1 **SSS Algorithm**

Following the discussion of summarization system we have to implement Sentence Scoring Summarization algorithm as mathematical basis for sentence scoring summarization method. Since the key part of summarization lies on sentence scoring and sentence selection followed by coherence, the quality of the generated summary can be ensured from the scoring function. The sentence scoring can be performed by term frequency: It takes into account only the frequency of a term inside the document:

TFi,j= number of occurrences of term in document j.

Document length: It is logical to assume that terms appear more frequently in bigger files, so if a term is relatively more frequent in a short than in a big file, then it is more important. To incorporate document length in the weighting formula we define:

DLj= total number of term occurrences.

This can be generalized to the average length of a document:

NDLj=DLj/ Average DL of all document

3.1.2. **CWE Algorithm**

The CWE (Content Word Extraction) algorithm extracts content words of the text document. Once the content words are separated we make dictionaries of noun, verb and named entities.

3.1.3. **HS-CFGs Algorithm**

The heart of the Headline Synthesis algorithm lies on HS-CFGs (Headline Synthesis using Context Free Grammar) algorithm.

Input:

Let l1, l2, l3 be length of the dictionaries representing noun, verb and named entities respectively. Let the nouns ,verbs, and named entities are already extracted out from the content words ,and let they are stored at dictionaries noun, verb and performers irrespectively.

Output:

Suitable headline for the text input.

Method:

Steps

[1]If l2>0 Add_production S -> NP,VP

[2]Else Add_production S -> NP

[3]Set NP->'the',N

[4]If l3>0 Do Steps 5 to 9

[5] Repeat Steps 6 to 9

[6] Add_production N ->item

[7] In performers_list

[9] Until l3>0

[10] Else Do steps 11 to 14

[11] Repeat Steps 12 to 14

[12] Add_production N ->item

[13] In noun

[14] Until l1>0

[15]If l2>0 Do Steps 16

[16] Add_production VP ->V ,'the' N

[5]Repeat Steps 15 to 17

[16] Add_production V ->item in verb

[17] Until l3>0

[19]Initialize rules

[19]Set expansion list Do step 20 to 23

[20]If the starting rule was in set of rules, then

[21]Grab one possible expansion

[22] For element in random_expansion Do

[23] Expand the element

[24]Else Do step 25 to 23

[25]If the rule wasn't found, then

[26] It's a terminal: Simply append the string to the expansion

[27]For every word in expansion

[28] If the word is repeating than

[29] Eliminate repeating word

[30]Output Headline

4. Manual Evaluation Technique

Manual Evaluation is simple set up in which the machine generated headline is evaluated manually. For a number of documents machine generated headline is compared against human generated headline. A suitable s core(mark)is assigned for both(human generated machine generated) the headline. The quality of headline generation system can be analyzed using suitable graphical method(bar graph or line graph).

5. Evaluation using vector space analysis

A vector space search involves converting documents into vectors. Each dimension within the vectors represents a term. If a document contains that term then the value within the vector is greater than zero. In this method both machine Generated and deviation of Human Generated headlines are converted into document vectors. A plot is performed by:

$\cos A = (t1*t2)/(t12) * (t22)$

6. Conclusion

The headline generation system is very successful for scientific and technical document, but less powerful for poetic language. Poetic language .the reason is that the poetic language contains less number of content words than that of scientific document. The size of document has also an impact on the generated headline. The machine generated title has also depends on how effectively document is summarized. The extraction of content words can also influence the headline. The Grammar used for sentence generation can influence the accuracy of the headline. The advantage of this headline generation system is that it does not require any learning to machine and the generated title depends entirely on input text file. Improved methods for keyword extraction and novel way for generating accurate sentences from input words will make the system powerful

7. References

[1] Automated Natural Language Headline Generation Using Discriminative Machine Learning Models Akshay Kishore Gattani B.E.(Honors). Birla Institute of Technology and Science Pilani (India) 2004.

[2] Automatic Text Summarization using a Machine Learning Approach .Joel Larocca Neto, Alex A. Freitas, Celso A. A. Kaestner .Pontifical Catholic University of Parana (PUCPR) Rua Imaculada Conceicao, 1155.

[3] Bengali Text Summarization By Sentence Extraction. Kamal Sarkar. Computer Science EngineeringDepartment Jadavpur University Kolkata 700 032 India.

[4] Challenges and Trends of Automatic Text Summarization. Oi Mean Foong1 , Alan Oxley1 , Suziah Sulaiman. Universiti Teknologi Petronas,Malaysi.

[5] Improved Algorithms For Keyword Extraction and Headline Generation From Unstructured Text.Amit Kumar Mondal and Dipak Kumar Maji. Department of Computer Science and Engineering Indian Institut of Technology, Kanpur Kanpur, India 208016.

[6] Natural Language Processing with Python.Steven Bird, Ewan Klein, and Edward Loper.

[7] Optimizing Machine Learning Approach Based on Fuzzy Logic in Text Summarization. Farshad Kyoomarsi, Hamid Khosravi Esfandiar Eslami, and Pooya Khosravyan Dehkordy. Islamic Azad University(Shahrekord branch) International Center for Science High Technology Environmental Sciences ,University of Shahid Bahonar Kerman Shahid Bahonar University of Kerman, The center of Excellence for Fuzzy system and applications.

[8] Sentence Extraction Based Single Document Summarization. Jagadeesh J, Prasad Pingali, Vasudeva Varma Workshop on Document Summarization, 19th and 20th March, 2005, IIIT Allahabad Report No: IIIT/TR/2008/97.

Krishnaprasad P have native place is at Palakkad, Kerala. He is doing his degree in B.tech Information Technology at Government .Engineering College Palakkad, Kerala, under the University of Calicut. His research interest includes areas such as NLP and Networking. Currently he is active at studying performance of TCP over wireless sensor network using NS-3.