Fuzzy Implementation of proposed hybrid algorithm using Topic Tilling & C99 for Text Segmentation

Mr. Mohit Lahori M.E Student
Computer Science & Engineering Deptt. ,
Chitkara University, Baddi , H.P India.

Dr. Shaily Jain Associate Prof.
Computer Science & Engineering Deptt.,
Chitkara University, Baddi , H.P India.

***Abstract* --The World Wide Web is a vast warehouse of information through which people are connected and are provided access to millions of resources via the internet. Although this information is represented in many types such as images, videos, etc., free-form text is by far the most common. A wide variety of topics constituting hundreds of millions of documents are available on the net. With the growing number of these documents, the need for an efficient and effective method to access the information also grows. Text and discourse processing techniques such as text classification, text segmentation and text summarization have been the ground work for organizing such information. Text segmentation, in particular, has been a very useful technique for segmenting unstructured text and building topic and subtopic structure, which can then be used to facilitate indexing and retrieval. In this paper we are going to present the hybrid of Text tilling and C99. In text tilling segmentation is performed in a hierarchical way. In c99 segmentation is done on a document in horizontal manner in blocks in a sequential way. In hybrid fuzzy logic is implemented on matrices of c99 & Topic Tilling to perform segmentation.**

Keywords -- Text segmentation, C99, Topic Tilling

I. INTRODUCTION

The aim of text segmentation is to divide a document into a group of segments which are coherent about a single topic. The task was carried out by taking into account the problems in information retrieval, summarization and language modeling, where smaller, coherent segments in a document are desired. The coherently segmented text makes way for efficient querying, analysis, and its usage. For example in information retrieval documents which are segmented topically results in short relevant text segments that corresponds to users query directly instead of scenario in which the user has to examine long documents to find the object of his/her interest. Topically segmented documents also produce a better summary than a number of segments which constitute a document.

Text segmentation can be understood as the segmentation of texts into topically similar units. It means viewing the text as a sequence of subtopics. A subtopic change marks a new segment. To find the sub topical structure of a text is main challenge for a Text Segmentation algorithm. There are two main approaches used for doing the aforesaid: lexical cohesion based approach and feature based approach. Lexical cohesion based approaches have a dependency on the inclination of topic units to be together. Approaches to measure such type of cohesion can be still further divided into two sub categories: Similarity based approaches where patterns of syntactic repetitions are used to indicate cohesion, and lexical chaining based approaches where other aspects of lexical cohesion (like relationships between terms) are also analysed. The second main category in text segmentation is that of feature based approaches where features like cue phrases, full detect boundaries proper nouns and named detect boundaries proper nouns and named entities are used to between topics.

II. MOTIVATION OR BACKGROUND

Morris and Hirst in 1991 [1] first proposed the notion of lexical chains to chain semantically related words (in fact synonyms) together via thesaurus. The chains are constructed out of selected terms in the document

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

and they represent the lexical cohesive structure of a text.

Hearst in 1994 [2] introduced TextTiling (TT) algorithm. This algorithm is a simple, domain-independent technique that assigns a score to each topic boundary candidate (inter-sentence gap) based on a cosine similarity measure between chunks of words appearing to the left and right of the candidate. Topic boundaries are placed at the locations of valleys in this measure, and are then adjusted to coincide with known paragraph boundaries. TT is straightforward to implement, and does not require extensive training on labelled data. It is designed to identify the subtopics within a single text and not to find breaks between consecutive documents. It segments the text at paragraph level.

The first hierarchical text segmentation algorithm was proposed by authors in [3] in 1997 using the cosine similarity and agglomerative clustering approaches. Different sources of information can be used to complement the lexical similarity.In this , evidence involving cue phrases and part-of-speech patterns can be executed on, using previously-trained decision trees, to help the lexical similarity function. Another research direction is table-of-content production.The main task here, and a major research topic, is identification of topics, or titles, for the segments. Finally, while the comparison with the TextTiling algorithm and the human judges is promising, a methodical evaluation of additional texts is required.

Reynar in 1998 [4] developed a method called dot plotting. This is a graphically motivated text segmentation technique. In this a similarity matrix is obtained from the text and then plotted on a graph. Dense regions on the graph corresponds to tight regions of topic similarity and are used to determine how the topic segments are distributed. This algorithm can be further enhanced for detecting both chapter boundaries in theworks of literature and story boundaries in Spanish news broadcasts. Certain suggestions & improvements i.e ways to make information retrieval, language modeling and various natural language processing algorithms efficient and can be exploited using the topic segmentation.

In [3] & [5],authors first introduced C99algorithm in 2000. This algorithm in fact uses a matrix-based ranking and a clustering approach in order to relate to the maximum similar textual units and to cluster groups of consecutive units into segments. Both TopicTilling and C99 characterize textual units by the words they contain.

Eisenstein in 2009 [6] created a hierarchical Bayesian algorithm based on LDA. In natural language processing, latent Dirichlet allocation (LDA) is a generative model that makes possible a sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.For topic modelling we use a generative model. This generative probabilistic model uses a training corpus of documents to create document-topic and topic-word distributations.

Martin Riedl and Chris Biemann in 2012 [13] formed TopicTiling algorithm. This algorithm is not based on words, but on the last topic IDs assigned by the Bayesian Inference method of LDA. This increases sparsity since the word space is reduced to a topic space of much lower dimension.
.

A. Detail of c99 [11][5] :

The main task of text segmentation is to get data from an image.C99 is in fact a text segmentation algorithm which segments a document in horizontal manner in blocks and segmentation is implemented in a sequential way. The basic unit for C99 is a block which is a group of words. Choi introduced the algorithm called C99 that uses two methods equally and together i.e a matrix-based numbering and a clustering method in order to obtain the most equal and similar textual units. Similar to the previous introduced algorithms, C99 uses words.The topic based version of the C99 algorithm, called C99 LDA, and divides the input text into some minimal blocks on the sentence boundaries. A similarity matrix S [m x m] has been calculated, in which m denotes the number of Units (sentences). Every unit element S [i,j] is computed using the concept of cosine similarity formulas between unit i and j. For these computations, each unit i is depicted as a T-dimensional Vector, where T denotes the actual number of topics selected for the topic model. Each single element T[k] of this particular vector contains the number of times topic ID k exists in unit i. Next, a rank matrix R is computed to enhance the contrast of S: Every single element R[i,j] contains the number of neighbours of S[i,j] that have smaller similarity scores then S[i,j] itself. This step makes an increase possible in the contrast between regions in comparison to original matrix S. In a concluding step, we have used a top-down hierarchical clustering algorithm to split the document into m segments. This algorithm gets started with a single whole document considered as one segment and splits it off into certain segments until the stop criteria are reached, e.g. the number of segments or a similarity

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

threshold. At this, the ranking matrix is split at indices i, j that maximize the inside density function D=$\sum_{k=1}^{m}$ sum of ranks/area within segments k.

B. Details of topic tilling [11][13]

Topic Tilling is a text segmentation algorithm. In Topic tilling a document is used in the form of a hierarchical distribution of topics and hence does segmentation in a hierarchical way. The basic unit in this algorithm is a topic which can be a single word or a set of words. Each single unit is given a topic id, based upon which the text segmentation is executed. A lot of research has been done on English literatures. Basically C99 (presented by choi) and Topic Tiling are used in many numbers to segment text in long literature documents but only a few have suggested to segment text from documents of images.In this section Text Segmentation algorithm called TopicTiling is introduced and discussed which is based on Text Tiling, but is in fact conceptually simpler. Topic Tiling considers a sentence is its smallest basic unit of text. Between each such position of p between two ad joint sentences, a coherence score cp is calculated. To compute the coherence score, we use the topic IDs assigned to the words by inference: Assuming an LDA model with T topics, each block is represented as a T-dimensional vector. The element of each such vector contains the frequency of the topic ID and obtained from the respective block. The coherence score is computed by cosine similarity for each adjacent "topic vector". Values close to zero indicate certain relatedness between two adjacent blocks, whereas values close to one denote a substantial connectivity.
Next, the coherence scores are tabulated to trace the local minima. These minima then are used as possible segmentation boundaries. But rather using the cp values, a depth score dp is calculated for each minimum [3]. In comparison to Topic Tiling, Text Tiling calculates the depth score for each such position and then searches for the local maxima. The depth score measures the deepness of a minimum by looking at the highest coherence scores on the left and on the right and is computed using the following formula:dp = 1/2 _ (hl(p) − cp + hr(p) − cp).The function hl(p) loops to the left as long as the score gets increased and returns the highest coherence score value. The same is done, looping in the other direction with the hr (p) function. If the number of such segments n is given as input, the n highest depth scores are used as segment boundaries. Otherwise, a threshold is applied. This threshold predicts a segmentation if the depth score is larger than $\mu - \sigma/2$, with μ being the mean and σ being the standard variation calculated on the depth scores. The algorithm runtime is in fact linear in the number of possible segmentation points, i.e. the number of sentences: for each segmentation point, the two adjacent blocks are sampled separately and combined into the coherence score.

III. EXPERIMENTAL SETUP:

The performance matrices to evaluate comparison are precision & recall.

1) Precision: Comparison of data in image file to the data which has been extracted by the algorithm used.

Formula is:-

P=Number of correctly system detected boundaries/Total number of system generated boundaries

2) Recall : Number of times the loop is iterated to get precise value.

Formula is:-
R=Number of correctly system detected boundaries/Total number of real boundaries.

1. file1.mat it is cricket.mat file.
Some text related to concept of cricket match.For humans the text appears as "UMPIRE SAID ITS A NO BALL DECISION WAS OK"

2. file2.mat it is dictionary.mat file.
Text of the concept morning walk. For humans the text appears as "LOT OF PEOPLE GO TO PARK IN THE MORNING IT IS GOOD FOR HEALTH "

3. file3.mat it is DTB1.mat file.
Text of the concept criminal case. For humans the text appears as "CRIMINAL CASE WAS REGISTERED AGAINST SANJAY DUTT"

4. file4.mat it is DTB2.mat file.
Text of the concept criminal case. For humans the text appears as "TOTAL SECURITY IN PC IS DONE BY ANTI VIRUS SO PROTECT PC"

5. file5.mat it is DTB3.mat file.
Text of the concept morning walk. For humans the text appears as "LOT OF PEOPLE GO TO PARK IN THE MORNING IT IS GOOD FOR HEALTH "

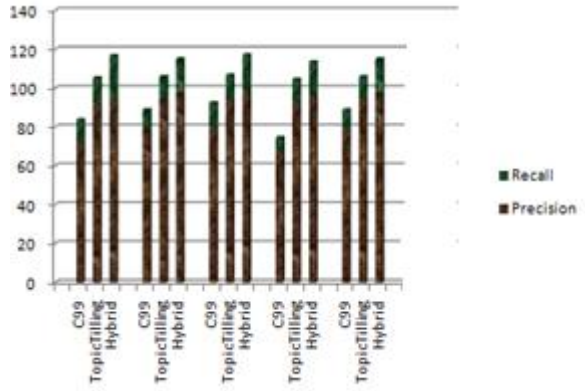
6. file6.mat it is football.mat file.
Text of the concept football match. For humans the text appears as "HE KICKED THE FOOTBALL AND DID GOAL"

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

IV. PROPOSED/METHODOLOGY

Explanation of fuzzy technique: Fundamentally Fuzzy logic is a well established technique in computer science. It gives us ways to deal with vagueness and uncertainty. Fuzzy logic starts with a set of rules. The fuzzy concept converts these rules to their mathematical equivalents. The gives more accurate & efficient way a system work e.g. in a computer.In the methodology & algorithm the two inputs to the fuzzy logic are the matrices generated by C99 & TopicTilling functions. To provide more accurate & efficient system design i.e. to give better results Fuzzy set (matrix1,matrix2) function in matlab has been used.matrix1 represent the boundary detected by c99 algorithm & matrix2 represent the boundary detected by TopicTilling algorithm. And Fuzzy set function finds the similarity of the two boundaries. Hence it is had made text segmentation efficient.

V. RESULT ANALYSIS



Hybrid is better in performance as compared to C99 & TopicTilling.In hybrid both C99 & TopicTilling are used together.In hybrid boundary identification of both the algorithms are evaluated together to get the resultant boundary of any text element ,which is a more accurate result.The evaluated parameters i.e precision & recall of both C99 & TopicTilling is less as compared to hybrid.

Data set/Data Inpu	Technique	Precision	Recall
file1.mat	C99	72.549	11.4802
file1.mat	TopicTilling	92.4242	13.2461
file1.mat	Hybrid	95.122	21.981
file2.mat	C99	79.4118	9.6194
file2.mat	TopicTilling	95.1807	10.9856
file2.mat	Hybrid	98.1818	17.124
file3.mat	C99	80.3571	12.3852
file3.mat	TopicTilling	94.3636	12.7276
file3.mat	Hybrid	97.9167	19.566
file4.mat	C99	67.6923	7.1834
file4.mat	TopicTilling	93.75	11.0938
file4.mat	Hybrid	96.1538	17.5296
file5.mat	C99	79.4118	9.6194
file5.mat	TopicTilling	95.1807	10.9856
file5.mat	Hybrid	98.1818	17.124

Fig1: Showing comparison of C99 ,TopicTilling & Hybrid

VI. CONCLUSION & FUTURE WORK

In this paper we have presented text segmentation and it has been the ground work for organizing such information. Text segmentation, in particular, has been a very useful technique for segmenting unstructured text and building topic and subtopic structure ,a hybrid algorithm has been implemented in the system with fuzzy logic. In future improvements can be done to make the concept more efficient.

VII. REFERENCES

[1]Morris, J., and G. Hirst. "Lexical cohesion computed by thesaurus relations as an indicator of the structure of text" *in Computational Linguistics* 17(1): 21–48. 1991.

[2]Marti A. Hearst, "Multi-paragraph segmentation of expository text" in Proceedings of *the 32nd Annual*

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET

Meeting of the Association for Computational Linguistics, LasCruces, New Mexico, USA, June 1994.

[3]Yaakov Yaari, "Segmentation of expository texts by hierarchical agglomerative clustering" in Proceedings of *RANLP'97,*1997.

[4]Jeffrey C. Reynar, "Topic segmentation: Algorithms and applications" PhD thesis, Computer and Information Science, University of Pennsylvania, 1998.

[5]Freddy Choi."Advances in domain independent linear text segmentation". In Proceedings of *NAACL-00,* pages 26–33,2000.

[6]Jacob Eisenstein "Hierarchical text segmentation from multi-scale lexical cohesion". In Proceedings of *NAACL09* 2009.

[7]Michael A. El-Shayeb, Samhaa R. El-Beltagy and Ahmed Rafea "Comparative Analysis of Different text Segmentation Algorithms on Arabic News Stories," in proceedings of *Information Reuse and Integration-IRI*, pp.441-446,2007.

[8]R. Barzilay and M. Ethadad,"Using Lexical Chains for Text Summarization," presented at Intelligent Scalable Text Summarization Workshop, Madrid, Spain, 1997.

[9]DoughBeeferman, Adam Berger, and John Lafferty, "Statistical models for text segmentation," *Machine Learning*, vol. 34, pp. 177 - 210, 1999.

[10]Teo Yung Kiat, "Linear and Hierarchical Text Segmentation Using Product Partition Models," Doctoral dissertation, Master Thesis, Department of Computer Science, School of Computing, National University of Singapore, 2004,2005.

[11]Martin Riedl, Chris Beimann, "Topic Tiling: A Text Segmentation Algorithm based on LDA" in proceedings of the *Association for Computational Linguistics*, Student research workshop, pp. 31-36, 2012.

[12] Qi Sun, Runxin Li,DingshengLuo, and Xihong Wu, "Text segmentation with LDA-based Fisher Kernel," in proceedings of the 46th Annual Meeting of *the Association for Computational Linguistics on Human Language Technologies*, pp. 269-272, 2008.

[13] Martin Riedl and Chris Biemann, "How text segmentation gain from topic models" in the proceedings of the annual conference of the North American chapter of *the Association for Computational Linguistics on Human Language Technologies*,Montreal, Canada, 2012.

[14] Sukhpreet Kaur, Kamaljeet Kaur Mangat, "comparative analysis of C99 and topictiling text segementation algorithms" in *International Journal of Research in Engineering and* Volume: 02 Issue: 09, Sep-2013.

ISSN: 2278 – 1323 All Rights Reserved © 2015 IJARCET