Main slides & sub-slides

1. Aim

a. Problem Statement

b. Brief description of work

2. Introduction

a. Previous work

b. Intuition behind the algo

3. Algorithm

a. Neural Network

b. SVD

c.Graphs

4. Dataset

a. Describe official dataset DUC (year, tasks, no. of documents)

b. Metric (Rougue score)

c. Automated tools used (like MEAD,UMich tools etc.)

5. Results

a. Rougue Scores obtained for different parts of dataset

b. Comparison with graphs,bar-graphs (variation of parameters)

c. Use of pie –charts to explain the effect

6. Conclusion/Future work

a. Based on findings

b. how they fulfilled the aim

c. Future work with more varied possibilities for exploration

7. References & Acknowledgements

a. Teachers (Bhatia Sir)

b. Papers referenced

c. Sites (wikipedia links etc.)

Content for slides

1.Aim

Problem Statement:

In this project we present a hybrid model for extractive text summarization based on neural Networks , singular value decomposition and graphs to generate automatic summary for the input document(s) by the user.

Brief description:

i) Our model uses both the supervised and the unsupervised settings, each reinforcing each other. We then combine the scores obtained from each step, by weighing each score (before adding them) with parameters calculated in an experimental setup to generate the final summary by selecting the top ranked sentences.

ii)Neural nets help the model in recognizing the important features (keywords, sentence position, sentence length) inherent in summary sentences and can be used by the user to train the summarizer to generate summaries of his choice(shorter sentences/thematic sentences etc.)

iii)To rank a sentence in the unsupervised setting we consider it’s value in two contexts: local (paragraph level) and global (document level). Singular value decomposition helps us in obtaining the importance of sentences in a global setting while graphs help us do so in a local context where we introduce a modified version of the longest path algorithm to implement the heuristic.

iv.) The weights are calculated by conducting experiments on the DUC 2004 data. Finally, we test the model on DUC 2004 data set where we obtain results comparable to the state of the art for our model.

2. Introduction

a.) Previous work

With the advent of technology, communication,mass media and social networking the quantity of text corpus has been growing at an ever increasing rate. Consequently, the problem of text summarization has received great attention from the research community all around the world.

Two major approaches were suggested in the past, i.e., extractive (Radev et al., 2004; Li et al.,2006) and abstractive

summarization (Dejong,1978).

i.)The abstractive approaches typically need to understand and paraphrase the salient information across the document.

ii.)In extractive approach the summarizer selects sentences from the passage that cover significant concepts in the documents. As a result these approaches tend to be more practical.

Early research on extractive summarization is based on heuristic features of the sentences such as their position in

the text, the overall frequency of the words they contain, or some key phrases indicating the importance of the sentences (Baxendale, 1958; Edmundson, 1969; Luhn, 1958). A commonly used measure to assess the importance of the words in a sentence is the inverse document frequency, or idf, which is defined by the formula (Sparck-Jones, 1972):

idfi = log\_N/ni ..(1)

where N is the total number of the documents in a

collection, and ni is the number of documents in which

word i occurs.

More sophisticated research on studying the contextual features that included some permutation of surface,content, relevance and event features. Surface features are related to extrinsic aspects of a sentence. **(**Kam-Fai Wong\*, Mingli Wu, Wenjie Li,2008) propose a model based on a supervised and semi-supervised setting to study the importance of several contextual features in sentence ranking.

A pioneering work in extractive summarization has been done by erkaan and radev(2004).In their work they consider the salience of a sentence by applying a Pagerank like algorithm that recursively computes the centrality of

sentences.As a prior step the relationship between sentences is calculated by the cosine similarity between sentence vectors in which each dimension corresponds to the term frequency and inverse document frequency product of each word in the bag of words in the documents.

Intuition behind Algo

Diversify the sentence ranking process unlike work done previously that focus on a single technique like sentence salience or centroidal sentences.

Using an experimental setup to obtain the the optimal weight given to each factor depending on required length of the summary helps in generating high quality summary based oon empirical testing.

Use of neural nets helps in extracting key features and eliminating the redundant ones,hence providing a user oriented approach.

Use of SVD helps in ranking sentences with their relevance in a global setting i.e. at the document level,and the matrix structure(arranged in descending order) helps in selecting important sentences from the most dominant topics in the documents.

Use of graphs helps us in ranking sentences with their relevance in a local setting (at the paragraph level),by using a modified version of the longest path algorithm.Using stongly connected components, graphs also help us in selecting sentences from distinct topics thus avoiding redundancy by selecting repetitively from same topics.

3.Algorithm

a.Neural Networks

(Refer PPt and ask mittal for content)

b.SVD

(done yourself)

c.Graphs

4.Dataset