Spoken Lecture Summarization (and Recommendation), using *only* Text Mining

Nirant Kasliwal (2012C6PS694P) & Saurabh Maurya (2012A7PS055P)

ABSTRACT

This project implements an improved approach for *spoken lecture summarizations*, in which random walk is performed on a graph constructed with automatically extracted key terms and probabilistic latent semantic analysis (PLSA).
Each sentence of the document is represented as a node of the graph and the edge between two nodes is weighted by the topical similarity between the two sentences. The basic idea is that sentences topically similar to more important sentences should be more important.
Each summary is re-scored to remove sentences which are very similar to sentences in the remaining summary corpus. This is done in an attempt to improve upon each summary's discrimination power in reference to the other summaries in the corpus.
A clustering mechanism based on soft cosine similarity is proposed to recommend similar summaries and lectures.

INTRODUCTION

Topic Motivation and Structure

In the Internet era, digital content over the network includes all the information and activities of human life. The most attractive form of network content is multimedia that may include speech. Such speech information usually tells the subjects, topics, and core concepts of the content. However, multimedia/spoken documents are just video/audio signals, usually much more difficult to retrieve and browse, because they cannot be easily displayed on the screen, and the user cannot simply skim through each of them from the beginning to the end. Hence, spoken document summarization becomes very important.

The project is divided into two logical structures. One is the automatic spoken lecture summarization and other is the recommendation. The primary focus of this report is on lecture summarization and the techniques involved in the same. Several heuristic based improvements and approximations have been made as and when the need to do the same was felt.

The improvements in the technique or idea from a text mining perspective have been reviewed in a separate section in addition to the mention as a part of the methodology.

A comparative review of the existing methods on similar, but not same dataset has been done.

General Approach

INSERT IMAGE HERE

DATASET

The primary corpus of the experiment was made using the TEDTalks Library. The library contains lectures of usually less than 20 minutes. Each of the transcript files is accessible using the (beta) TED API.

The initial dataset was a json corpus with detailed metadata about the timestamp of each sentence. This was cleaned to extract sentence by sentence of the transcript. The same has been used as the primary initial corpus throughout the experiment.

EXPERIMENT

Proposed Methodology

1. Sentence Scoring

1.1 PLSA PLSA has been widely used to analyze the semantics of documents based on a set of latent topics.

PLSA was used as the statistical measure for the purpose of this project.

1.2 Bigram and Trigram Scoring Each sentence is processed to find the possible bigrams and trigrams. The bigram and trigram score of the sentence is used to highlight correlation between frequently reoccurring phrases.

The score increases as the number of such words increases.

1.3 Part of Speech Noun phrases are extracted from each sentence. Noun phrases have the highest semantic value and represent the key topic of the sentence almost completely. Hence, other parts of speech have been assigned zero value.

2. Random Walk

As mentioned in the base paper, we formulate the sentence selection problem as random walk on a directed graph. Each sentence is a node and the edges between them are weighted by topical similarity. The basic idea is that a sentence similar to more important sentences should be more important. In this way all sentences in the document can be jointly considered more globally rather than individually.

A random walk over the graph allows for each visited node's sentence score to be re-evaluated. This rescoring is done on the basis of the similarity between the source sentence node and destination sentence node.

3. Variance Maximization

In a large corpus such as ours, the lecture summaries tend to lose its *discrimination power*. The entire corpus is flattened into one large document. Sentences which show high similarity to each other are eliminated from the new corpus. Each summary now has higher variance in comparison to the entire corpus than earlier. This increases the discrimination power of the summary corpus and the ability to distinguish each document more accurately.

Two methods - one as suggested by the 2011 work by Chen et al and the other as mentioned by Han et al. were used to find sentence similarity scores. The work by Han et al. is chosen because it has the best known performance.

4. Recommendation

The summaries were clustered on soft cosine similarity. The clustering is used to recommend the related summaries.

RESULTS AND DISCUSSION

In the experiments to be presented below, the summarization ratio was set to be 10% and 20% and compared. The key phrases (with more than one df word) automatically extracted were taken as individual terms in PLSA modeling and all following processes.

A few comparisons have been made: The first is use of PLSA as statistical input to the sentence score and summary without the same.

The second comparison is the summarization with rescoring from random walk over a graph and without the same.

The last comparison is by removing the sentence which are similar to other sentence in the corpus. This is done by removing sentence which were closer than 1 standard deviation to the variance.

Each of the above comparison was made by human inspection.

Statistical Measure

In all cases, PLSA scores outperformed the simple scoring without it. This is probably because the key term knowledge is very helpful, especially for manual transcriptions. This is probably because in manual transcriptions all key terms are correctly transcribed (although may be incorrectly extracted) so that the key-term-based statistical measures were much more accurately estimated.

Random Walk

The introduction of random walk led to shuffling of the sentence ranks without the same. The end results showed a range of changes varying between no change (where the summary

had no new sentence nor sentence reordering) to large changes (where the summary has new sentences after the ranking change). Most of the summaries showed medium to no change on visual inspection.

Corpus Wide Variance Maximization

This step of the pipeline relied on elimination of similar sentences across the corpus. It was observed that the use of the same sentence scoring as used for the random walk led to worse results than we would expect. The improvements sections highlights the remedial action taken to do the same.

IMPROVEMENTS

Corpus Wide Variance Maximization

During the random walk, similar sentences might receive high scores and be included in the final summary. This step in the pipeline hopes that to compensate for the bias that may have been introduced because of the same.

This method is specially useful in a situation where there are several lectures which have closely related topics. These lectures, hence exhibit high correlation and similarity between the extracted summary sentences.

Alternative Sentence Similarity Scoring

An alternative semantic text similarity system was proposed by Han et al. It was done as a part of the Semantic Textual Similarity Task 2013, hosted by ACM.

The proposed idea used a lexical similarity feature that combined POS tagging, LSA word similarity and WordNet knowledge. The paper makes the following concluding claims: "The first run, which achieved the best mean score out of all 89 submissions, used a simple term alignment algorithm augmented with two penalty metrics. The other two runs, ranked second and fourth out of all submissions, used support vector regression models based on a set of more than 50 additional features."

An alternative arrangement of the experiment used the Han similarity score to eliminate the sentence instead of the same score used earlier over random walk. The need to do was felt because elimination based on the same document specific scoring led to worse results. In some cases, the summary was reduced to less than three sentence with the scoring without WordNet knowledge.

CONCLUSION

Automated lecture transcript summarizartion is an increasingly important topic of interest. With strides in automated speech recognition, this can act as the last part of the pipeline to yield a completely automated summary of a recorded lecture. This would allow a viewer/student to quickly skim through the content.

This experiment demonstrated the utility of topic semantic analysis. Additionally, it also highlighted the fact that the similarity score of sentence is dependent on the context and knowledge against which it is measured.

ACKNOWLEDGMENTS

We wish to thank our instructor in charge Dr. Poonam Goyal. It was her mentorship, guidance and motivation that enabled us to strive well beyond our prior capabilities in the domain.

Additionally, we would like to thank our friends who are helping us to prepare the human summaries.