Generating slides from HTML pages

Arun JVS
201001079

Rohit Girdhar
201001047

Sudheer Kumar
201001149

Abstract—**The abstract goes here.**

I. Introduction

This demo file is intended to serve as a "starter file" for IEEE conference papers produced under LaTeX using IEEEtran.cls version 1.7 and later. I wish you the best of success. [1]

II. Related Work

There has been limited work that addresses the problem of automatic slides generation. Previous works include approaches like section-wise summarization [?], alignment methods for matching document regions with presentation regions [?], extracting topics and itemized elaborations from tagged documents [?]. Many of these systems use multiple ideas from Information Retrieval, such as TF-IDF weights, query expansion, query-specific summarization, POS-tagging, etc.. Some of them are purely generative, while others learn models from an existing corpus of document-presentation pairs. These systems are tuned to work with different input formats such as PDF, XML, LaTeX and PPT, each of which preserve a varying amount of semantic and structural information about the original text. In the following two subsections, we review two papers [?],[?] which particularly deal with technical papers.

A. SlidesGen

The first work we review is by Sravanthi et al. [?]. They propose an novel framework for automatic generation of presentation slides for technical papers. They take LaTeX documents of research papers as input, and return the presentation slides. Their method depends on the assumption that by and large, conference papers have a similar structure: an abstract, followed by sections that can be broadly classified as introduction, related work, actual work (model), experiments, conclusion/results and bibliography. A slide in their system contains a title and some bulleted points that are important in that section. They evaluate the system by surveying the response of people who use their system.

Their system is divided into multiple stages. The first stage is pre-processing stage. In this step, the LaTeX documents are converted into XML using a public domain converter LaTeXML.

Next, the generate what they call ”configuration file”, which contains configuration parameters for each section (since each section has a different point of view and writing style). This involves categorization of the section and extraction of key phrases from the section which are then stored in the configuration file. For example, a section with large number of cite tags, or with title containing words such as ”related work” or ”literature survey” is categorized as related works section.

The next step is of extracting key phrases. Most research papers have associated key phrases that can be help categorize the content in paper, and contain important concepts introduced in the paper. They are mostly related to model and experiments section, and can be used to summarize the same. So, the keywords given at the beginning are added to keyphrases for those sections in the configuration file. Also, few other phrases as the names of subsections etc are also added to the configuration file.

Next, they use QueSTS summarizer to summarize the model, experiment and conclusions section. QueSTS represents the text as a Integrated Graph where sentence is a node, and edge exists between 2 sentences if the sentences are similar. The edge weight is defined as the cosine similarity between the 2 sentences (above a given minimum threshold) Given the keyphrases computed previously, they are tokenized and a centrality based (to the query phrase) node weight is computed for the node corresponding to each token. Thus, they get query specific node weights and query independent edge weights.

From the above graph, they construct a Contextual Tree for each term q and node r. All the trees at node r are merged into a Summary Graph at r. The SGraphs generated from each node are ranked using a scoring model, and the one with best rank is returned as the summary.

The final step of this procedure is slide generation from the XML and configuration file. The introduction slides are generated by comparing introduction sentence with the abstract using consine similarity and ones with high similarity are placed in the slides. They give a similar method for generating slides from related work section as well. From model and experiments section, the slides are generated using the above QueSTS mechanism. Similarly, conclusion slides are generated using comparison with keywords such as ”proposed”, ”concluded” etc.

Another important aspect in slides is graphics. Graphics are added along with sentences that either refer to it or is present along with it.

The paper finally discusses the issues in alignment of sentences and generation of slides. They also evaluate their slides manually with multiple users, taking their feedback. Most users gave a satisfaction level of more than 8/10 to the slides generated.

B.

References

[1] George D. Greenwade. The Comprehensive Tex Archive Network (CTAN). *TUGBoat*, 14(3):342–351, 1993.