Introduction

Surveys are helpful in quick knowledge acquisition in a given topic. However, in rapidly advancing and specialized fields like Natural Language Processing (NLP) and Artificial Intelligence (AI), surveys quickly become outdated or simply do not exist. Currently, automatically generated surveys in these two domains are not yet reliable enough for general use. One aspect of automatically generated surveys that has room to improve on is its structure. Current systems that attempt this typically disregard the hierarchical structure within documents and instead focus on a one dimensional ordering of extracted text.

In this paper, we propose a method where the hierarchical structures of input documents are parsed into trees which are then combined to generate a hierarchical structure for a survey. Initial tests of the system yield promising results that show both potential of using hierarchical structure in survey generation and that certain aspects of the system still have room to improve.

Materials and Methods

Five topics were selected from the AAN TutorialBank corpus. For each topic, three to six documents were selected as inputs to the system.

The hierarchical structure (headings and subheadings) were extracted into a tree and each node was converted into a vector representation using modified tf-idf. Due to the short length of headings, the tf portion also takes ancestor nodes of a given heading into account. The idf portion was trained on papers from the ACL Anthology in order to include relevant terms.

Each document tree was then combined to form a final tree that forms the structure of a survey. Cosine similarity was used to decide node merges.

The final generated structure can be displayed using a pre-order depth-first traversal of the generated tree.

Conclusion

The initial results are promising and suggest that the general hierarchical approach to generating the structure of surveys may be effective. The term similarity portion appears to work well, while other components will benefit from further testing and development. To aid this, a larger corpus of well-parsed documents in different topics will be helpful. Overall, the current implementation is able to produce a reasonable result with certain sections of the generated tree that show potential for future development.

Acknowledgement

Thank you to Alex Fabbri and Professor Dragomir Radev for advising me. I would also like to thank Michael Yasunaga and Ryan Lim for helping to evaluate my results.