Automatic text summarization using Rhetorical Structure Theory

Madhav Narayan, Department of Computer Science, narayan@cs.utexas.edu
Elisa Ferracane, Department of Linguistics, elisa@ferracane.com

TEXAS
The University of Texas at Austin

Background

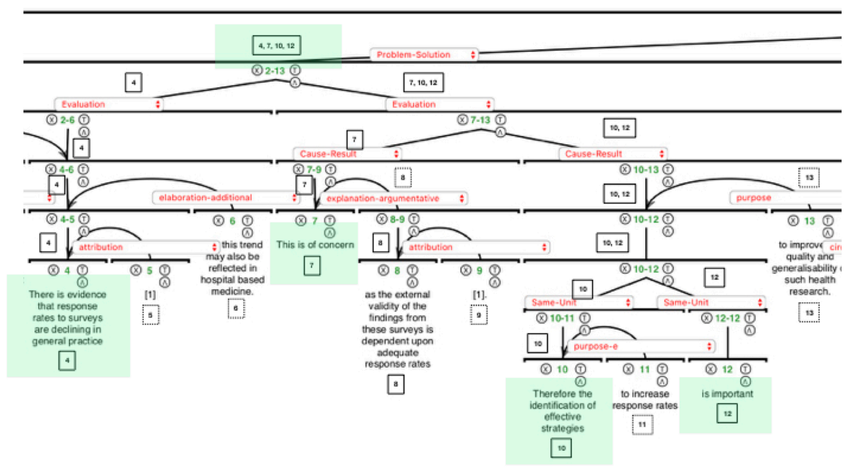
- As the amount of digital text increases, manually reading and extracting important information from these texts becomes highly time-consuming.
- Automatic text summarization enables us to efficiently summarize large amounts of text and easily identify the important, relevant parts.
- One approach to summarizing text is *discourse processing*, which examines the logical structure of text and the relations between its different parts. These relations are formalized by *Rhetorical Structure Theory* (RST), which can be used to identify the relative importance of different sentences and phrases in a text.

Research Question

How effective are (manual) annotations using Rhetorical Structure Theory in producing *extractive summaries* of medical documents?

Methods and Algorithm

- Given a document labelled with RST relations, we construct a *discourse tree* that represents the hierarchical relations between textual units.
- Next, we perform a *promotion process* on the tree, "promoting" textual units that we consider important up through the tree.
- This process yields an ordering on the textual units in the tree based on the relative importance ("salience") of textual units.
- We then use this ordering of textual units to construct a summary for the document.



An example of a discourse tree that illustrates how salient textual units are promoted to the top of the tree. This discourse tree yields part of the summary given below. Starting with nuclei at the leaf nodes, textual units move up through the tree as they remain nuclei for their parent nodes. The higher a node makes it up the tree, the more important it may be considered when creating a summary of the document.

Response rates to postal questionnaires are falling and this threatens the external validity of survey findings. We wanted to establish whether the incentive of being entered into a prize draw to win a personal digital assistant (PDA) would increase the response rate for a national survey of consultant obstetricians and gynaecologists.

Actual summary

There is evidence that response rates to surveys are declining in general practice. This is of concern. Therefore the identification of effective strategies is important. We hypothesised that entry into a prize draw would increase response to a survey of practising obstetricians and gynaecologists.

Generated summary

Comparison of a generated summary with the actual (abstractive) summary of the Background section in a medical paper. Text highlighted in green represents text that is semantically represented in the generated summary; text in red represents details not retrieved in the generated summary; and text in orange represents text that is paraphrased or potentially implied in the generated summary.

Ten patients dropped out of the study from each treatment arm. There was a significant, marked improvement in HAM-D and MADRS scores in each group by the treatment endpoint. There was no significant difference between PB and sertraline groups on either HAM-D or MADRS at any visit. The response rate was 90\% with PB and 92\% with sertraline. The remission rate was 70\% with PB and 75\% with sertraline. All laboratory parameters were within normal limits in all patients. There were no serious adverse events.

Actual summary

B was as effective as sertraline in attenuating depression rating scores on the HAM-D and MADRS. The recommended dosing range of PB is 75-300 mg/day. It is noteworthy that the mean endpoint dose was 250 mg/day. PB was tolerated as well as or better than sertraline.

Generated summary

An example of a summary that fails to retrieve the important information in a document when compared to the actual (abstractive) summary for the document.

Discussion

- Our initial conclusion is that this approach to creating summaries is effective and intuitive, but still susceptible to shortcomings.
  - In long documents, salient units are often unable to be promoted to the top of the discourse tree.
  - Irregularities in annotations can lead to some salient units being missed.
- Going forward, we would like to consider multiple discourse structures for each document, and determine ("learn") optimal characteristics of discourse structures, which can lead to better summaries.

Acknowledgments

We would like to thank Dr. Katrin Erk for the opportunity to work on this project, and for her guidance and support. We would also like to thank Titan Page for his work on annotating the medical documents with RST information. This project was done as part of the IE Pre-Graduate School Internship program.

References

[1] Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, *8*(3), 243-281.

[2] Marcu, D. (1998). Improving summarization through rhetorical parsing tuning. In *Sixth Workshop on Very Large Corpora*.