UNIVERSITY OF MASSACHUSETTS AMHERST

Content Summarization of a Book using its Index

TL;DR - An attempt to create a TL;DR

Pankaj Bhambhani

Department of Computer Science

CS 585 Natural Language Processing

Email: `pbhambhani@cs.umass.edu`

Motivation

Explosion of digital information all over the globe, new generation finds books pretty long to read. Generating summaries as an attempt to bridge the gap. Index of a book provides key areas to look.

Scope

Focused on **centrality** based summarization algorithms, **extractive** summarization and on books with a back-of-the-book index

Dataset

Dataset of 10 ebooks from **Project Glutenberg** and **UMass Amherst Libraries**, Reference summaries for evaluation taken from **wikisummaries.org**. Average length of a book is around **88750 words** while that of summary is around **2450 words**.

Approach/Method

Extract content text, pre-process the data extract sentences based on the index words, and summarize the filtered sentence list.

Evaluation

Evaluation based on ROUGE-1 Metric, used for summarization and machine translation.

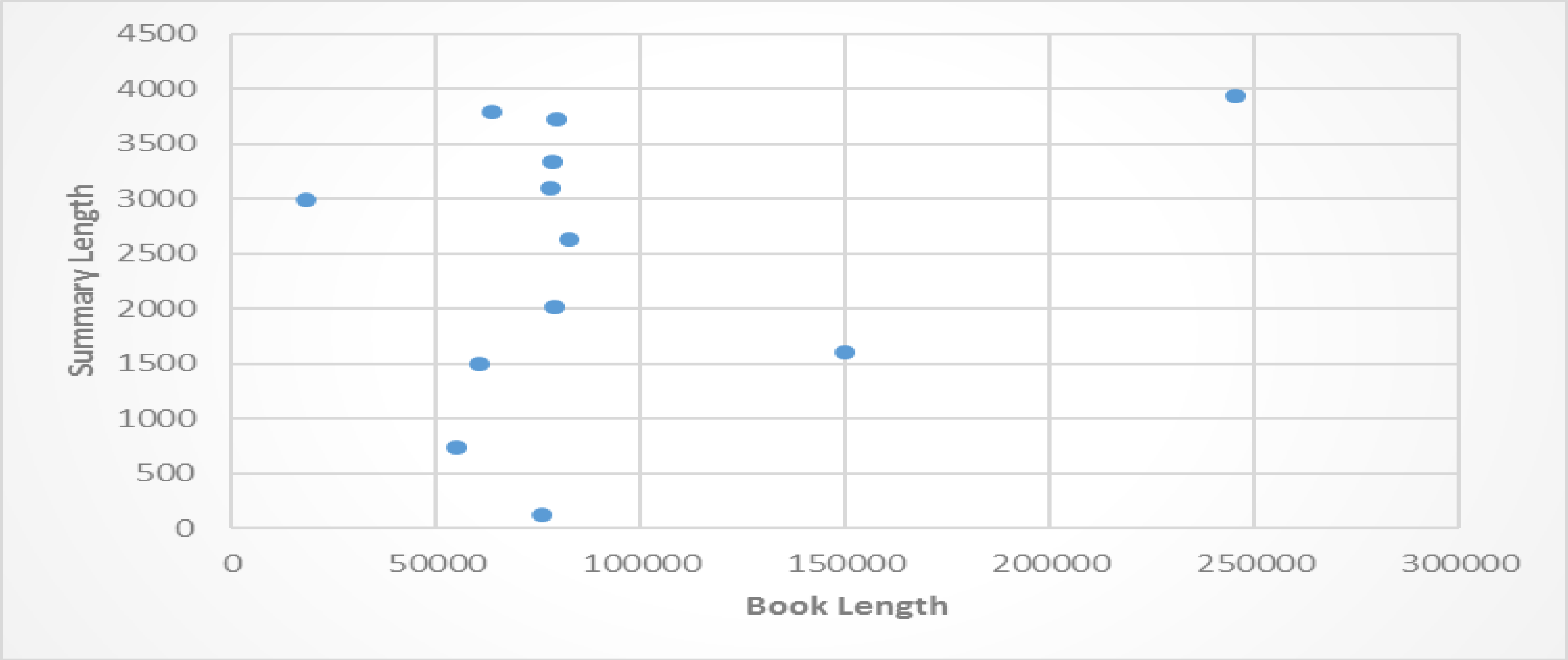

Figure 1: Plot of Book Length vs Summary length

Summary Type	Avg R	Avg P	Avg F
Full Text	0.68	0.27	0.38
Index Filtered (LexRank)	**0.59**	**0.34**	**0.41**
Index Filtered (TextRank)	**0.67**	**0.28**	**0.39**
Baseline	0.53	0.41	0.44

Table 1: Evaluation comparison for LexRank and TextRank

$$imc(x,y) = \frac{\sum_{w \in x,y} tf_{w,x}\, tf_{w,y}\, (idf_w)^2}{\sqrt{\sum_{x_i \in x}(tf_{x_i,x} idf_{x_i})^2}\sqrt{\sum_{x_i \in x}(tf_{x_i,x} idf_{x_i})^2}}$$

$$p(u) = \frac{d}{N} + (1-d) \sum_{v \in adj[u]} \frac{imc(u,v)}{\sum_{z \in adj[v]} imc(z,v)}\, p(v)$$

Figure 2: Centraility equations for LexRank

Conclusions/ Future Research

Index based summarization achieves higher F-score than using the full-text, computationally less expensive but doesn't do as good as baseline summary. In general, most document summarizer algorithms find it tough to beat baseline summaries. Better pre-processing, use of index page-numbers can improve summarization.

References

[1] Günes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479, 2004.

[2] Rada Mihalcea and Hakan Ceylan. Explorations in automatic book summarization. In *EMNLP-CoNLL*, volume 7, pages 380–389, 2007.

Acknowledgments - Professor Brendan O'Connor and TAs Haw-Shiuan Chang and Abe Handler.