Tweet Summarization Proposal

Marcus Hughes, jmh3

1 Overview

Tweets often correlate with movements and events, e.g #LoveWins trended after the Supreme Court's legalization of same-sex marriage in the United States in 2015 or #Ferguson trended after the social unrest following a police shooting in 2014. With a large enough sample of tweets one can theoreticaly enumerate, summarize, and understand the views of the public response to issues. One might also perform argumentation mining resulting in a large dataset of content-rich tweets, i.e. tweets with thesis and some piece of support, with evidence portions identified. However, there are still too many tweets and pieces of evidence to digest, hence the need for summarization.

I intend to explore the background literature on tweet summarization. In addition, I will implement cutting-edge summarization algorithms for tweets. Finally, I hope to develop a guideline of the strengths and weaknesses of each approach.

This project will serve many purposes. First, it familiarizes me with existing work in summarization before beginning thesis work in earnest. In addition, implementing promising summarization algorithms allows for a second study during my thesis: determining whether argument mining techniques in addition to summarization provide measurable advantage in characterizing public opinion through tweets and producing complete summaries over pure summarization approaches. If an advantage is measured, I can explore how to optimize it.

2 Prior Work

Multi-document summarization is divided into two classes: extractive, approaches that *extract* important phrases from the documents, and abstractive (or generative), approaches that *generate* an organic summary beyond copy and pasting important phrases. There is extensive work in extractive summarization including many papers on summarization of tweets. Abstractive summarization is much more difficult and has not yet been as fruitful. **SumBasic**, an algorithm developed by [3], is an example of an extractive approach that exploits frequency information. Its simplicity leads it to often be used as a baseline. This was later expanded into **SumFocus** resulting in better performance and more topical summaries [8]. **SumBasic** and **SumFocus** are just one approach to summarization algorithms among a multitude, e.g. graph representations [1, 7], clustering methods [2], machine learning [4], neural networks [5], Wordnet features [6].

3 Intended Approach

I will begin by utilizing previously mined tweets on Planned Parenthood (#StandWithPP and #DefundPP). However, I would like to eventually mine my own tweets (or at least understand the infrastructure utilized to retrieve these tweets.) The approach in this project is summarization with two goals: review and replication of existing approaches. If time allows (and inspiration provides), I will propose the foundations of a novel approach.

I will read a few papers (at minimum one) daily on summarization approaches and add those to a reference table throughout this project.

3.1 Timeline

- **4/27/17** Project beginning
- **4/28/17** Procure experimental dataset
- **4/29/17 - 4/30/17** Implement SumBasic
- **5/2/17** Complete a first draft of a table of previous research (clearly not exhaustive)
- **5/3/17** Decide what approach(es) I would like to replicate
- **5/4/17 - 5/9/17** Work on replicating an approach
- **5/9/17** Mid-way project presentation
- **5/9/17 - 5/15/17** Continue replicating approach(es)
- **5/16/17** Insure sufficient documentation and notes that any unfinished work could be resumed
- **5/17/17-5/18/17** Write final paper
- **5/19/17** Final report and code submission

References

[1] Erkan and D Radev. Lexrank: Graph-based lexical centrality as salience in text.

[2] Manjula, Sarvar Begum, and D Venkata Swetha Ramana. Extracting summary from documents using k-mean clustering algorithm. 2(8), 2013.

[3] Aniv Nenkova and Lucy Venderwende. The impact of frequency on summarization. In *Microsoft Research, Redmond, Washington Tech. Rep. MSR-TR 2005-101*, 2005.

[4] Joel Larocca Neto, Alex A Freitas, and Celso A A Kaestner. Automatic Text Summarization using a Machine Learning Approach.

[5] G Padmapriya and K Duraiswamy. An approach for text summarization using deep learning algorithm. *Journal of Computer Science*, 10(1):1–9, 2014.

[6] Alok Ranjan Pal, Projjwal Kumar Maiti, and Diganta Saha. An Approach To Automatic Text Summarization Using Simplified Lesk Algorithm And Wordnet. *International Journal of Control Theory and Computer Modeling (IJCTCM)*, 3(45), 2013.

[7] K Patil and P Bradzil. Sumgraph: Text summarization using centrality in the pathfinder network. 2007.

[8] Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing and Management*, 43(6).