Text Summarization as an Assistive Technology

Fahmida Hamid
Department of Computer Science
University of North Texas
Denton, Texas
FahmidaHamid@my.unt.edu

Paul Tarau
Department of Computer Science
University of North Texas
Denton, Texas
tarau@cs.unt.edu

ABSTRACT

Automated *text summarization* can be applied as an assistive tool for people with vision deficiency as well as with language understanding or attention deficit disorders. In this paper, we introduce an unsupervised graph based ranking model for text summarization. Our model builds a graph by collecting words, and their lexical relationships from the document. We apply a handful of available semantic information (definition, sentimental polarity) of words to enhance edge-weights (interconnectivity) between nodes (words). After applying a polarity based ranking algorithm over the graph we collect a subset of high-ranked and low-ranked words, name those as keywords. We, then, extract sentences that possess a higher rank defined by the rank vector of keywords. Sentences extracted in this manner correlate with each other and express the summary of the document quite successfully. Summaries formed by our model can appease readers with vision difficulties while keep them updated.

Categories and Subject Descriptors

Natural Language Processing [**Miscellaneous**]; Assistive Technology [**Augmentative and Alternative Communication**]

General Terms

Natural Language Processing, Automatic Text Summarization

Keywords

text summarization, rank, bias, graph

1. INTRODUCTION

Assistive Technology (AT) is the field of study concerned with providing devices and techniques to augment the abilities of a disable person. One of the largest groups benefitted by assistive technologies are people with visual impairment or vision loss. Vision loss results from either disease, trauma, or degenerative conditions that cannot be rectified entirely by medication. People suffering from vision loss need special assistance from the society to adapt themselves to daily life. We address text summarization (an area of *Natural Language Processing*) to assist these people. The goal of automatic summarization is to rephrase the most important content of the source in a condensed form according to the user or the application. One can easily guess how great it will be to have some tools for generating the essence of a news or an article from the perspective of a person with reading deficiencies. An automatic general purpose text summarization tool would be of immense utility not only to people with vision difficulties as well as language understanding or attention disorders. but also to a person / system that handles large archives of documents. For instance, one needs to organize a digital library with a collection of several million books. A reliable document summarizer is possibly the best option that can help grasp the topics discussed in the document in order to be categorized. Hence, we easily can visualize the applications and importance of a text summarizer.

A summary is a text that is produced out of one or more texts, that contains the same information of the original text, and that is no longer than half of the original text [2]. In order to summarize, one needs to determine the relationship between sentences in a document. Words are the building blocks of a sentence. Dissecting lexical syntax could help one understand relationships amongst words in a sentence quite clearly. On the other side, Sentences, are related to each other semantically. It is hard to determine their relations for several issues like, anaphora resolution, word sense disambiguation and so on. Hence there is no explicit rule to select a subset of sentences to represent the summary of the document.

There has been a comparatively new trend in *Natural Language Processing* that uses graph based ranking algorithms [3] to process texts and extract keywords or sentences quite successfully. In this work, we have applied a graph based ranking algorithm [4] that works on weighted graphs. This approach originally was proposed for finding trustworthiness on trust-based networks (social networks, peer-to-peer networks etc). We adapted their idea on text-graphs, which works with not only opinion-biased but also non-biased texts.

2. RELATED WORK

PETRA '14, May 27 - 30 2014, Island of Rhodes, Greece.

ACM 978-1-4503-2746-6/14/05...$15.00.

Distinctive works have been done for text summarization. Gong & Liu [1] used two techniques (rank sentence relevance, latent semantic analysis) to find the summary. Both of their approaches are basically greedy methods who select highly ranked sentences with less redundancy. In order to determine the relevancy of a word, they used *term-frequency (tf)* as a feature. In order to determine the sentence of interest, one can utilize the word statistics. The typical frequency of occurrence of words averaged across the length of a document is an intuitive way of collecting *keywords*.

Graph based ranking algorithms (HITS, PageRank etc.) have successfully been used for citation analysis, social networks and analysis of link-structure of the World Wide Web. Mihalcea and Tarau [3] have applied a similar line of thinking to lexical or semantic graphs extracted from natural language documents, which results in a graph based ranking model that can be applied to a variety of natural language processing applications. The authors have discussed about applying their *TextRank* algorithms in directed or undirected, weighted or un-weighted graphs.

A document can be treated as a collection of opinions around some topics. The summary of a document should contain the major opinions about major topics. This idea tempted us to represent a text-graph, with weighted interconnections (edges) and entities, like words (nodes) with different polarity. We adopted Mishra and Bhattacharya's network model [4] to represent the text-graph. Their model represents a network of nodes based on the *trust-scores* they earn: Each node possesses two properties:- bias and prestige. **Bias** reflects the expected weight of an outgoing edge, whereas **prestige** reflects the expected weight of an incoming edge.

Formally, let $G = (V, E)$ be a directed graph with set of vertices V and set of edges E, where $E \subset (V \times V)$. For a given vertex v, let $d^o(v)$ be the set of vertices pointed by v (successor), and $d^i(v)$ be the set of vertices that points to v (predecessor). Then, *bias*, *prestige* and other related terms can be generated using the following equations.

$$bias(i) = \frac{1}{2|d^o(i)|} \sum_{j \in d^o(i)} (w_{ij} - prestige(j)) \quad (1a)$$

$$prestige(j) = \frac{1}{|d^i(j)|} \sum_{k \in d^i(j)} w_{kj}(1 - X_{kj}) \quad (1b)$$

$$X_{kj} = max\{0, bias(k) \times sign(w_{kj})\} \quad (1c)$$

$$w'_{kj} = w_{kj}(1 - X_{kj}) \quad (1d)$$

The above formulas of bias and prestige enable the model to handle two main design problems in a trust based network, handling negative weights and distinguishing edge with *zero* weights.

3. TEXT AS GRAPH

To apply the graph based ranking algorithms to natural language texts, we need to represent text as a graph. Though we will work with 'sentence's, we decided to start from 'word's. Our algorithm performs the following steps:-

- Collect Signature Words.
- Create nodes and edges.
- Add edges (Update weights).
- Apply formula [1] over the graph until the *rank* value converges.
- Create a set of *keywords* by selecting top (2/3)rd of high ranked, (1/3)rd of low ranked words.
- Create an *weight vector* from the rank-values of *keywords*.
- Use this vector to rank the sentences.
- Select top (1/3)rd of ranked sentences to present the *summary*.

3.1 Signature Words

- Decompose each sentence into words & remove stop-words
- Collect words who has parts of speech labeled as {'noun', 'adjective', 'verb' , 'adverb'} & place them in candidate list of **signature** words
- Find proper **definition** of each signature word
- Find **sentimental polarity** of each signature word, & set this as **bias** value of the word.

Example 01: *This country cannot afford to be materially rich and spiritually poor.*

Word	PoS	Polarity	Definition
country	n	0.0	a politically organized body of people under a single government
afford	v	0.0	be able to spare or give up
materially	r	0.0	with respect to material aspects
rich	a	0.0	possessing material wealth
spiritually	r	0.125	in a spiritual manner
poor	a	0.0	deserving or inciting pity

Table 1: Words & its Entities

3.2 Nodes & Edges: From a single sentence

Let, x, y be two words residing in the same sentence, and $|position_x - position_y| < windowSize$; then we create distinct nodes (if not already exists) for x and y, and define their relations (edges) by either of the rules:

1. If $parts\_of\_speech(x) = \{verb\}$, add $edge(x, y)$.
2. If $parts\_of\_speech(x) \cup parts\_of\_speech(y) \subset \{noun, adjective, adverb\}$, then add $edge(x, y)$ and $edge(y, x)$.

Finally, we add $weight(x, y) = (total\_Edges)^{-1}$ to all the existing edges. The sentence graph for Example 01(with $windowSize = 4$) is shown in figure 1.

3.3 Update Edge Weights

Let x and y be two different words from two different sentences (or in the same sentence, $|position_x - position_y| \geq windowSize$) in the original document. We use their *definition* (which is human annotated data, available in WordNet) to determine *similarity* between them. If there is an existing edge between x and y, we adjust it by putting extra *weight*. Otherwise, we add a new edge. The simplest way to determine the *similarity* between x and y is to find the total common words in their definition over the length of x's (or largest) definition. We do not add any edge if the *similarity* is *zero*.

This phase helps to relate semantically closer words in the document.

3.4 Keyword Extraction

Once the graph is built, all the nodes and edges are set to some bias and weight values, we add a real value (can

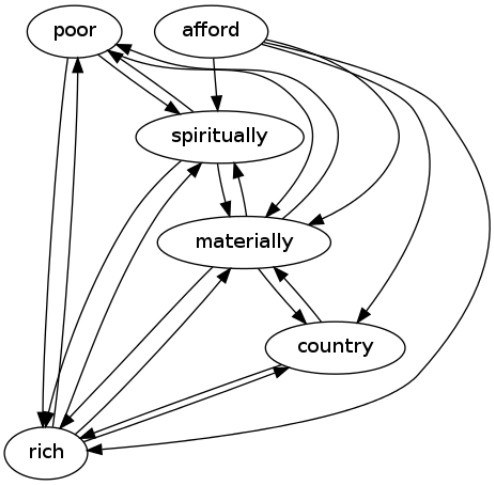

Figure 1: Sentence Graph

Word	Definition	Similarity
program	a series of steps to be carried out or goals to be accomplished	
start	take the first step or steps in carrying out an action	0.076923
program alone	<same> without any others being included or involved	0.00
society	an extended social group having a distinctive culture and economic organization	
community	a group of people living in a particular area	0.090909

Table 2: Degree of Similarity

keyword	Rank	
student	0.0675132436788	high-ranked
right	0.0660165720559	
nation	0.041348048758	
year	0.036937558335	
urban	0.0343118064486	
country	0.0327797923094	
citizen	0.0326381753249	
shortage	0.0322712200123	
service	0.0322303624747	
child	0.0315760583231	
regulation	0.000174857202176	low-ranked
doctor	0.000146226571885	
race	0.000142663672242	
needle	0.000129017628287	
research	0.000127861244762	
counsel	0.000101960018603	
working	7.18129799462e-05	
capacity	5.59380076004e-05	
waste	4.95651368777e-05	
sense	3.25298427849e-05	

Table 3: A set of Keywords

sentence	weight
Today, an estimated 4 out of every 10 students in the 5th grade will not even finish high school - and that is a waste we cannot afford	6.73338399846e-05
Moreover, all our miracles of medical research will count for little if we cannot reverse the growing nationwide shortage of doctors, dentists, and nurses, and the widespread shortages of nursing homes and modern urban hospital facilities	8.79322162439e-05

Table 4: Sample of top sentences

be chosen randomly) to every node as it's *rank*. This way, there is no discrimination beforehand. Then we apply set of equations(1) several times (until the rank value converges) over the graph. For real time output, one can control the repetition using a threshold. Negative bias and weighted edges help us decide the *importance* of the words, by changing their rank values. Table 3 shows top 10 high ranked and top 10 low ranked keywords (for article at section 4), which are a subset of 38 final keywords that our program generated out of 299 signature words. The idea behind choosing low-ranked as well as high-ranked nodes is to provide importance to the *negative* opinions as well as *positive* ones.

3.5 Sentence Extraction

Our top (high and low) ranked signature (key) words define the weight vector for sentences. For this set of experiments, we have got best results with top $(2/3)rd$ high ranked and top $(1/3)rd$ low ranked keywords.

If $v = \{v_1, v_2, \ldots, v_k\}$ be the top k keywords with corresponding weight vector $w = \{w_{v_1}, w_{v_2}, \ldots, w_{v_k}\}$, then for a sentence s_j,

$$weight(s_j) = (\sum_{v_m \in s_j} w_{v_m})/(k \times |s_j|) \quad (2)$$

One can decide to select top y (we chose the top (1/3)rd) sentences to form candidate *summary* based on their *weight* vector. To avoid promoting long sentences, we are using length of the sentence as the normalization factor and divide the $weight(s_j)$ with the length of sentence s_j. Table 4 shows two top sentences out of selected 11 for article at section 4; words highlighted with blue represent keywords from *high-ranked* section and with green represent keywords from *low-ranked* section.

4. EVALUATION

We have applied our algorithm over fifty articles found in *nltk-database* which comprises stories, speeches, general discussions or debates. We also have tested the *keyword extraction* phase on two-hundred abstracts collected from *NIPS (Neural Information Processing System)*. In each case, our algorithm could successfully extract the major keywords and generate meaningful summaries.

Out of a set of tested files, we show an input file (Example Article, section 4) with 25 sentences, and our program generated summary file (Example Summary, section 5) with 11 sentences. Each sentence in the *summary* file contains the sentence itself and corresponding weight value attached at the end. We build the graph with words, and finally ap-

ply the word-weight (rank) vector to determine the rank of the sentences. For sentence extraction task, Mihalcea and Tarau[3] prepared the graph where the nodes were the sentences; then they ran the 'textrank' algorithm to find more important sentences. Both approaches work reasonably well; but with this polarity based model, we could express opinion biased texts more intuitively .

Example Article

Tax reduction alone, however, is not enough to strengthen our society, to provide opportunities for the four million Americans who are born every year, to improve the lives of 32 million Americans who live on the outskirts of poverty. The quality of American life must keep pace with the quantity of American goods. This country cannot afford to be materially rich and spiritually poor. Therefore, by holding down the budgetary cost of existing programs to keep within the limitations I have set, it is both possible and imperative to adopt other new measures that we cannot afford to postpone. These measures are based on a series of fundamental premises, grouped under four related headings:

First, we need to strengthen our Nation by investing in our youth: The future of any country which is dependent upon the will and wisdom of its citizens is damaged, and irreparably damaged, whenever any of its children is not educated to the full extent of his talent, from grade school through graduate school. Today, an estimated 4 out of every 10 students in the 5th grade will not even finish high school - and that is a waste we cannot afford. In addition, there is no reason why one million young Americans, out of school and out of work, should all remain unwanted and often untrained on our city streets when their energies can be put to good use. Finally, the overseas success of our Peace Corps volunteers, most of them young men and women carrying skills and ideas to needy people, suggests the merit of a similar corps serving our own community needs: in mental hospitals, on Indian reservations, in centers for the aged or for young delinquents, in schools for the illiterate or the handicapped. As the idealism of our youth has served world peace, so can it serve the domestic tranquility. Second, we need to strengthen our Nation by safeguarding its health: Our working men and women, instead of being forced to beg for help from public charity once they are old and ill, should start contributing now to their own retirement health program through the Social Security System. Moreover, all our miracles of medical research will count for little if we cannot reverse the growing nationwide shortage of doctors, dentists, and nurses, and the widespread shortages of nursing homes and modern urban hospital facilities. Merely to keep the present ratio of doctors and dentists from declining any further, we must over the next 10 years increase the capacity of our medical schools by 50 percent and our dental schools by 100 percent. Finally, and of deep concern, I believe that the abandonment of the mentally ill and the mentally retarded to the grim mercy of custodial institutions too often inflicts on them and on their families a needless cruelty which this Nation should not endure. The incidence of mental retardation in this country is three times as high as that of Sweden, for example - and that figure can and must be reduced. Third, we need to strengthen our Nation by protecting the basic rights of its citizens: The right to competent counsel must be assured to every man accused of crime in Federal court, regardless of his means. And the most precious and powerful right in the world, the right to vote in a free American election, must not be denied to any citizen on grounds of his race or color. I wish that all qualified Americans permitted to vote were willing to vote, but surely in this centennial year of Emancipation all those who are willing to vote should always be permitted. Fourth, we need to strengthen our Nation by making the best and the most economical use of its resources and facilities: Our economic health depends on healthy transportation arteries; and I believe the way to a more modern, economical choice of national transportation service is through increased competition and decreased regulation. Local mass transit, faring even worse, is as essential a community service as hospitals and highways. Nearly three-fourths of our citizens live in urban areas, which occupy only 2 percent of our land - and if local transit is to survive and relieve the congestion of these cities, it needs Federal stimulation and assistance. Next, this Government is in the storage and stockpile business to the melancholy tune of more than $16 billion. We must continue to support farm income, but we should not pile more farm surpluses on top of the $7.5 billion we already own. We must maintain a stockpile of strategic materials, but the $8.5 billion we have acquired - for reasons both good and bad - is much more than we need; and we should be empowered to dispose of the excess in ways which will not cause market disruption. Finally, our already overcrowded national parks and recreation areas will have twice as many visitors 10 years from now as they do today. If we do not plan today for the future growth of these and other great natural assets - not only parks and forests but wildlife and wilderness preserves, and water projects of all kinds - our children and their children will be poorer in every sense of the word.

5. CONCLUSION

Assistive Technology is impartially a broad area of research. Natural Language Processing (NLP) techniques can be applied for AT (specially augmentative and alternative communication) in a large variety of ways, for example, providing communicative assistance for frail people or individuals with severe vision impairments. In this work, we focus on a technique of summarizing texts of a single document. We believe techniques like this can be of great use to people with vision loss. We have shown a new model for single document summarization.

Our approach is domain independent and unsupervised. We have applied the proposed method over a various type of documents, for example, news articles, abstracts of scientific articles, historic speeches or even for some random articles collected from wikipedia. In each case, our algorithm generated a sensible summary. Our model works quite successfully for single document summarization.

Example Summary

■ This country cannot afford to be materially rich and spiritually poor. ■ 7.8420555764e-05 ■ These measures are based on a series of fundamental premises, grouped under four related headings: First, we need to strengthen our Nation by investing in our youth: The future of any country which is dependent upon the will and wisdom of its citizens is damaged, and irreparably damaged, whenever any of its children is not educated to the full extent of his talent, from grade school through graduate school. ■ 4.13368156469e-05 ■ Today, an estimated 4 out of every 10 students in the 5th grade will not even finish high school - and that is a waste we cannot afford. ■ 6.73338399846e-05 ■ Moreover, all our miracles of medical research will count for little if we cannot reverse the growing nationwide shortage of doctors, dentists, and nurses, and the widespread shortages of nursing homes and modern urban hospital facilities. ■ 8.79322162439e-05 ■ Third, we need to strengthen our Nation by protecting the basic rights of its citizens: The right to competent counsel must be assured to every man accused of crime in Federal court, regardless of his means. ■ 7.21905755844e-05 ■ And the most precious and powerful right in the world, the right to vote in a free American election, must not be denied to any citizen on grounds of his race or color. ■ 7.89624064217e-05 ■ Fourth, we need to strengthen our Nation by making the best and the most economical use of its resources and facilities: Our economic health depends on healthy transportation arteries; and I believe the way to a more modern, economical choice of national transportation service is through increased competition and decreased regulation. ■ 3.80563820614e-05 ■ Local mass transit, faring even worse, is as essential a community service as hospitals and highways. ■ 7.80544529572e-05 ■ Nearly three-fourths of our citizens live in urban areas, which occupy only 2 percent of our land - and if local transit is to survive and relieve the congestion of these cities, it needs Federal stimulation and assistance. ■ 5.41192050256e-05 ■ Finally, our already overcrowded national parks and recreation areas will have twice as many visitors 10 years from now as they do today. ■ 0.000125691097628 ■ If we do not plan today for the future growth of these and other great natural assets - not only parks and forests but wildlife and wilderness preserves, and water projects of all kinds - our children and their children will be poorer in every sense of the word.■ 3.71754772964e-05

As an extension to this work, we plan to combine some *rule-based* algorithm which will help us resolute anaphora between nouns and pronouns in the following sentences. We experimented with shorter articles where we applied *anaphora resolution* by hand and it performed much better on defining sentence connectivity and rank related words more precisely. We plan to extend this work and build a model that can generate summary not only by extracting sentences but also by rephrasing some from the original one.

6. REFERENCES

[1] Y. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25, New York, NY, USA, 2001. ACM.

[2] E. Hovy and C.-Y. Lin. Automated text summarization and the summarist system. In *Proceedings of a Workshop on Held at Baltimore, Maryland: October 13-15, 1998*, TIPSTER '98, pages 197–214, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.

[3] R. Mihalcea and P. Tarau. TextRank: Bringing order into texts. In *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*, July 2004.

[4] A. Mishra and A. Bhattacharya. Finding the bias and prestige of nodes in networks based on trust scores. In S. Srinivasan, K. Ramamritham, A. Kumar, M. P. Ravindra, E. Bertino, and R. Kumar, editors, *WWW*, pages 567–576. ACM, 2011.