An Insight into Automatic Ergonomic Text Summarization Approaches

Amar Nair Syed Ahmed Vishnu Vidyan
Department of Computer Science
University of Central Florida, Orlando, FL 32716
amarknair@knights.ucf.edu, pID: 3941593
syed@knights.ucf.edu, pID: SY3722427
vishnu4v5@knights.ucf.edu, pID: V3717832

Abstract

Summaries are an important tool for familiarizing oneself with a subject area. With the increase in the availability of online information, the practical need for automatic summarization has become important within the Natural Language Processing community. We demonstrate automated summary construction with a) Maximal Marginal Relevance (MMR) as the baseline approach with an aim to reduce redundancy while maintaining query relevance in re-ranking retrieved documents and b) the introduction of sentence salience to identify most important sentences in documents with degree-based approach called LexRank to compute centrality using the similarity graphs. We then evaluate the results of our approach on Document Understanding Conference (DUC) 2004 dataset using ROUGE toolkit and finally, discuss the challenges and pending problems that remain open.

1 Introduction

Summarization is one of the sought after challenges of natural language processing since it involves both the analysis of an existing text and the composition of a new one. The various techniques in summarization are also proven to be helpful in paraphrasing to check comprehension, identifying relevant information, consolidating similar pieces of information and increasing Information Retention. The main goal here is to produce a minimized version of documents that capture only the required words/sentences of the original document so as to provide its core meaning. As defined by Radev et al. (2002) (Radev et al., 2002), *"a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that"*.

Most successful summarization systems utilize extractive approaches that crop out and stitch together portions of the text to produce a condensed version. In automatic summarization, however, one often distinguishes between informative and indicative summaries, where informative summaries intend to make reading of source unnecessary, if possible. It is a well-researched field and involves two main categories: a) Text extraction which aims at identification of most relevant sentences in multi-document datasets using standard statistically based information retrieval techniques augmented with more or less shallow natural language processing and heuristics.(Jing and McKeown, 2000) And b) Text abstraction on the other hand aims at interpretation of the text semantically into a formal representation in order to parse the original text in a deep linguistic way and find new and more concise concepts to describe the text and then generate a new shorter text, an abstract, with the same information content.

Creating summaries that somehow resemble human summaries is very challenging, there is an empirical limit intrinsic to pure extraction (Genest et al., 2013), also summaries lack coherence and cohesion because they are merely excerpts from the text. Earliest instances of research on summarizing sci-

entific documents proposed paradigms for extracting salient sentences from text using features like word and phrase frequency, position in the text and key phrases. Various work published since then has concentrated on other domains, mostly on newswire data. Many approaches addressed the problem by building systems depending of the type of the required summary. In a paradigm more tuned to information retrieval (IR), one can also consider topic-driven summarization, that assumes that the summary content depends on the preference of the user and can be accessed via a query, making the final summary focused on a particular topic. The work by Das and Martins (Das and Martins, 2007) addresses contemporary methods in extractive summarization. One of the main problems in this field is distinguishing the more informative parts of a document from the less ones and requires tedious text-to-text generation involving identification of topic(s), selection of the most salient sentences from which information is to be derived and generation of sentences based on the information obtained from the salient parts of the text (Gatt and Reiter, 2009). Last but not the least, a very crucial issue that will certainly drive future research on summarization is evaluation. During the last fifteen years, many system evaluation competitions like TREC, DUC and MUC have created sets of training material and have established baselines for performance levels. However, a universal strategy to evaluate summarization systems is still absent.

2 Motivation and Problem Formulation

Relevance and non-redundancy are the two main characteristics of a good summary, the major concern of any summarization technique is effective representation of results. One such approach based on clustering was introduced by Yongzheng et al. (Radev et al., 2002) and the below formulation provides the intuition behind such approaches.

$$S_i = w_1 * C_i + w_2 * K_i + w_3 * Ti + w_4 * L_i$$

where S_i is the score of sentence i. C_i, K_i and T_i are the scores of the sentence i based on the number of cue words, keywords and title words it contains, respectively. L_i is the score of the sentence based on its location in the document. w_1, w_2, w_3 and w_4 are the weights for linear combination of the four scores.

Once the documents are clustered, sentence selection from within the cluster to form its summary is local to the documents in the cluster. But these techniques still have problems. The frequency distribution of the words in any document in fact follows a power law and there is both a lower cut-off and an upper cut-off on the frequency of words that should be used to determine the salience of sentences for a document (Luhn, 1958). When pre-processed data is used, the importance of the proposed methods are diminished. Moreover, in traditional summarization techniques, the importance of sentence is measured only by its location which provides very weak results in dynamic environment. Also, important or relevant information is usually spread across sentences, and extractive summaries cannot capture this and hence it fails to present conflicting information accurately.

As per Endress work (Endres-Niggemeyer et al., 2000), users prefer extractive summaries over the glossed abstractive summaries. One of the reasons being the former presents the information as-is by the author allowing the users to read between-the lines. The traditional extractive methods score sentences in order to generate summaries with the most important ones and the score in most of the methods is computed based on the type of document. The next section describes how these problems can be solved using maximal marginal relevance (MMR) measure and stochastic graph-based methods.

3 Our Approach

This section describes in detail the approaches we have followed to implement a baseline and a novel system for automatic generic text summarization. We start with Maximal Marginal Relevance method as the baseline approach, followed by the novel approach - LexRank. The goal in both the approaches in to obtain meaningful and promising summaries as per the task2 guidelines of DUC 2004. Once both the paradigms have been implemented, we test for performance evaluation using the ROUGE evaluation kit which has been described in the next sections of this paper.

3.1 Maximal Marginal Relevance

We use the concept of topic-driven summarization introduced by Carbonell and Goldstein (1998) (Witte et al., 2007) with the maximal marginal relevance (MMR) measure. The idea is to combine query relevance with information novelty; it may be applicable in several tasks ranging from text retrieval to topic-driven summarization. MMR simultaneously rewards relevant sentences and penalizes redundant ones by considering a linear combination of two similarity measures. Let Q be a query or user profile and R a ranked list of documents retrieved by a search engine. Consider an incremental procedure that selects documents, one at a time, and adds them to a set S. So let S be the set of already selected documents in a particular step, and R/S the set of yet unselected documents in R. For each candidate document $D_i \in R/S$, its marginal relevance $MR(D_i)$ is computed as:

$$MR(D_i) = \lambda Sim_1(D_i, Q) - (1-\lambda) \max_{D_i \in S} Sim_2(D_i, D_j)$$

where λ is a parameter lying in $[0, 1]$ that controls the relative importance given to relevance versus redundancy. Sim_1 and Sim_2 are two similarity measures; in the experiments both were set to the standard cosine similarity traditionally used in the vector space model,

$$Sim_1(x, y) = Sim2(x, y) = \frac{x.y}{|x|.|y|}$$

The document achieving the highest marginal relevance, $DMMR = argmax_{D_i \in R/S} MR(D_i)$, is then selected, i.e., added to S, and the procedure continues until a maximum number of documents are selected or a minimum relevance threshold is attained. As suggested, we first start with $\lambda \approx 0.3$ and then gradually increase the value to 0.7 in order to focus on the most relevant documents. If we move beyond single document summarization to document cluster summarization, where the summary must pool passages from different but possibly overlap ping documents, reducing redundancy becomes an even more significant problem. Within a topic, the document clusters are processed in chronological order and a summary generated for each cluster by arranging the high ranked sentences until the limit of 100 words is reached.

Sentences containing new information (i.e. that could not be inferred by any previously considered document) are selected to generate summary. However, this highly efficient approach and requires large linguistic resources. Witte et al., (Witte et al., 2007) propose a rule-based system based on fuzzy coreference cluster graphs. Again, this approach requires to manually write the sentence ranking scheme. Several strategies remaining on post-processing redundancy removal techniques have been suggested. Extracts constructed from history were used by Boudin et al., (Boudin and Torres-Moreno, 2007) to minimize historys redundancy. Lin et. al.,(Lin et al., 2007) have proposed a modified Maximal Marginal Relevance (MMR).

The MMR-passage selection method for summarization works better for longer documents (which typically contain more inherent passage redundancy across document sections such as abstract, introduction, conclusion, results, etc.). MMR is also extremely useful in extraction of passages from multiple documents about the same topics.

News stories contain much repetition of background information. Our preliminary results for multi-document summarization show that in the top 10 passages returned for news story collections in response to a query, there is significant repetition in content over the retrieved passages and the passages often contain duplicate or near replication in the sentences. MMR reduces or eliminates such redundancy.

3.2 LexRank

LexRank is a stochastic graph-based method of computing sentence importance for text summarization proposed by Erkan and Radev (Erkan and Radev, 2004a). This method represents each sentence as a node and computes the cosine-similarity between sentences to generate edges between them. The entire method depends on the fact that if the similarity between two sentences is above some predefined threshold then there is an edge between the two nodes otherwise there isn't. This method can outperform the centroid based methods and is quite insensitive to the noise in the data. As it is mentioned in the original paper, a collection of similar

documents can be viewed as a network of sentences. Some sentences would share only little information with the rest of the sentences. But there would be salient sentences which would be similar to a number of sentences in the cluster. This can be evaluated as more central sentences.

A cluster of documents can be viewed as a network of sentences that are related to each other. Some sentences are more similar to each other while some others may share only a little information with the rest of the sentences. We hypothesize that the sentences that are similar to many of the other sentences in a cluster are more central (or salient) to the topic. There are two points to clarify in this definition of centrality. First is how to define similarity between two sentences. Second is how to compute the overall centrality of a sentence given its similarity to other sentences (Erkan and Radev, 2004b).

In LexRank, we represent a cluster of documents by a cosine similarity matrix. Each entry in the matrix would be the similarity between a corresponding sentence pair. IDF modified cosine-similarity is calculated as follows:

$$IDF_{mod}(x,y) =$$

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

where tfw_s is the number of occurrences of the word w in sentence s. We calculate TF-IDF values for all the words in the cluster by finding out term frequency for a word in a given sentence, IDF value for a given word and IDF value for all the words in a document cluster. With this we would be able to calculate the similarity score between two sentences. We designed a cosine threshold of 0.1 for this similarity (Brandow et al., 1995). From experiments, it generated a better summary compared to other values of the threshold we tried. It is considered that too low threshold may take sentences with very low similarities and too high threshold may miss out similar sentences in the document cluster. This intra sentence cosine similarities defines the graph structure and the weights between the nodes (which are sentences here).

We use the famous PageRank formula to calculate the saliency of the sentences. Following is the matrix form of the page rank equation:

$$p = [dU + (1-d)B]^T p$$

where centrality vector p corresponds to the stationary distributions of the similarity matrix B. U is a square matrix with all elements being equal to $1/N$.

The transition kernel $[dU+(1d)B]$ of the resulting Markov chain is a mixture of two kernels U and B. To compute the stationary distribution, we use an iterative algorithm called Power Method as explained in the original paper. The only difference between the original PageRank and the lexical PageRank is that the prior one is directed graph and the latter is undirected one (Mihalcea et al., 2004). But the computations remain the same.

4 Dataset

The document understanding conference (DUC) was the main forum providing benchmarks for researchers working on document summarization. The tasks in DUC evolved from single-document summarization to multi-document summarization. We evaluated our implementation discussed in the previous section on DUC 2004 consisted of 500 newspaper and newswire articles in which the most important evaluation measure used is coverage, intended to capture how well summarizers perform content selection and not addressing issues such as readability and other text qualities of the summaries. There are 2 generic summarization tasks (Tasks 2, 4a, and 4b) in DUC 2004 which are appropriate for the purpose of testing our approaches. Task 2 involves summarization of 50 TDT English clusters. In this task we perform text clustering. We then extract the sentences by selecting the top ranked sentences, from the top ranked segments from the top ranked clusters. These sentences are then arranged in a chronological order, by sorting them with respect to the time stamps of the documents they are extracted from, generating a summary ($<=$ 665 bytes).

5 Experiments

As mentioned in the previous section, we have used a cluster of documents pertaining to 50 topics from the DUC2004 data corpus. Along with these human

summaries have been used for the performance evaluation of the two paradigms which have been implemented on python (version 2.7). We have used PyRouge (version 0.1.0) and ROUGE toolkit (version 1.5.5) on Ubuntu 14.04 for performance evaluations.

5.1 Maximal Marginal Relevance

Since the extractive summarization which we are to implement is a generic form as opposed to a query based approach, we had to generate a query sentence before beginning the summary extraction. The query is generated by ranking all the words in the document cluster in the order of their TF-IDF score and then selecting from them a particular number of words with the highest scores. We had set the query length to 10 words as increasing beyond this did not seem to improve the results notably. Also for measuring the similarity for pairs of sentences we calculated the cosine similarity. Once the query was generated the best sentence from the document cluster is extracted by measuring and ranking the similarity scores for the sentences in the cluster with reference to the query. Once this is complete the MMR scores of the rest of the sentences in the cluster can be calculated and then added incrementally. Each time the marginally relevant sentence with maximum score is added to the summary it is removed from the document cluster so that it is not considered during the next iteration. In our experiments, the implemented paradigm was tested with varied values of lambda and tested from $\lambda = 0.3$ to $\lambda = 0.7$, the results have been reported in the Table1 in section 5.3.

5.2 LexRank

For this part of the experiments we had to summarize 50 topics (indexed from d30001 to d31050) which contained news articles reported by major news channels and agencies. The LexRank was designed to generate a 100-word summary for each topic using the text files as input. We have designed the lexRank so that it generates a 5 sentence output summary. From practice it was understood that it is generating a 100 words or more summary on an average. The entire program was developed in python. The LexRank summarization is stored as LexRank.py in the project folder. The program is taking 4-5 seconds for running for a single topic. Hence in total the program is taking a little more than 4 minutes to run. In a cluster of related documents, many of the sentences are expected to be somewhat similar to each other since they are all about the same topic. This is shown in Figure 1.

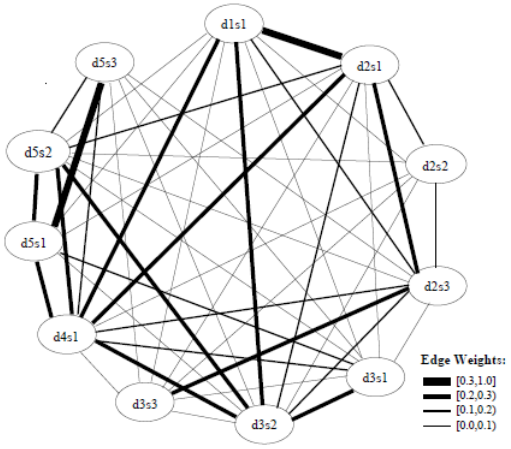

Figure 1: Weighted cosine similarity graph for the cluster in Table 1

5.3 Results

ROUGE was originally developed by Chin-Yew Lin (Mihalcea et al., 2004), a researcher at Microsoft Research. ROUGE as an evaluation tool generally measures the n-gram overlap between the system and human summaries. It has multiple metrics. ROUGE-1, ROUGE-2, and ROUGE-SU4 are the most commonly used metrics. They respectively measure the unigram (one word), bigram (two consecutive words), unigram and skip-bigram (two words with a gap of up to 4 words in the middle) overlap between the system and human summaries. The higher the score is, the better the system performance is. ROUGE-2 has traditionally been considered to correlate well with human judgments on news document summarization evaluation. Nowadays it is a typical practice to report all three metrics together. The implementation is based on Perl but we used the Python interface package named PyROUGE since the exposed interface is simple. We used the following parameters and have also explained the significance of each of them:

-n 4 -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -a -x -l 100.

- **"-n 4"** compute Rouge-n up to max-ngram

S #	ID	Text
1	d1s1	Iraqi Vice President Taha Yassin Ramadan announced today, Sunday, that Iraq refuses to back down from its decision to stop cooperating with disarmament inspectors before its demands are met.
2	d2s1	Iraqi Vice president Taha Yassin Ramadan announced today, Thursday, that Iraq rejects cooperating with the United Nations except on the issue of lifting the blockade imposed upon it since the year 1990.
3	d2s2	Ramadan told reporters in Baghdad that Iraq cannot deal positively with whoever represents the Security Council unless there was a clear stance on the issue of lifting the blockade off of it.
4	d2s3	Baghdad had decided late last October to completely cease cooperating with the inspectors of the United Nations Special Commission (UNSCOM), in charge of disarming Iraqs weapons, and whose work became very limited since the fifth of August, and announced it will not resume its cooperation with the Commission even if it were subjected to a military operation.
5	d3s1	The Russian Foreign Minister, Igor Ivanov, warned today, Wednesday against using force against Iraq, which will destroy, according to him, seven years of difficult diplomatic work and will complicate the regional situation in the area.
6	d3s2	Ivanov contended that carrying out air strikes against Iraq, who refuses to cooperate with the United Nations inspectors, will end the tremendous work achieved by the international group during the past seven years and will complicate the situation in the region.
7	d3s3	Nevertheless, Ivanov stressed that Baghdad must resume working with the Special Commission in charge of disarming the Iraqi weapons of mass destruction (UNSCOM).
8	d4s1	The Special Representative of the United Nations Secretary-General in Baghdad, Prakash Shah, announced today, Wednesday, after meeting with the Iraqi Deputy Prime Minister Tariq Aziz, that Iraq refuses to back down from its decision to cut off cooperation with the disarmament inspectors.
9	d5s1	British Prime Minister Tony Blair said today, Sunday, that the crisis between the international community and Iraq did not end and that Britain is still ready, prepared, and able to strike Iraq.
10	d5s2	In a gathering with the press held at the Prime Ministers office, Blair contended that the crisis with Iraq will not end until Iraq has absolutely and unconditionally respected its commitments towards the United Nations.
11	d5s3	A spokesman for Tony Blair had indicated that the British Prime Minister gave permission to British Air Force Tornado planes stationed in Kuwait to join the aerial bombardment against Iraq.

Table 1: Document cluster d1003t from DUC 2004

- **"-e"** specify the data folder comes with ROUGE
- **"-m"** use stemming
- **"-2 -4 -u"** use unigram and skip-bigram with distance up to 4 (aka ROUGE-SU4)
- **"-l 100"** use first 100 words of summary for evaluation
- **"-c 95"** confidence level
- **"-r 1000"** an option used in resampling
- **"-x"** do not calculate Rouge-L

ROUGE-2 recall with stemming and stop-words not removed provides the best agreement with manual evaluations. We also compute ROUGE-1 recall, which is the measure with highest recall of ability to identify the better summary in a pair, and ROUGE-4 recall, which is the measure with highest precision of ability to identify the better summary in a pair. We have reported the recall, precision and f-score for our implementation below.

As mention in the section 5.1 we tested the MMR scores for incremental values of λ from 0.3 to 0.7 to get an understanding of the variation in the summaries being generated and it was found that λ values closer to zero makes the system focus on answering the respective queries while λ values closer to 1 makes the system focus on reducing the redundancy on the summaries being generated. We chose to select λ value that would concentrate on both the mentioned parameters equally and hence have reported the scores for $\lambda = 0.5$

6 Overlap between summaries

Using the ROUGE scores we have been able to compare the efficiency of MMR and LexRank summarization systems. However, in order to analyse the degree of similarity of summaries generated by these systems we studied the overlap of the generated summaries in terms of words and sentences.

6.1 Sentence Level Comparison

Since the summaries generated by both the systems are extractive, selecting sentences directly from the document cluster, it should be easy to measure the degree of sentence overlap of summaries produced for the same set of inputs by these systems. For this purpose, we have used the Jaccard coefficient to compute the degree of sentence overlap. The Jaccard coefficient score can be calculated as:

$$J(A_S, B_S) = \frac{|A_S \cap B_S|}{|A_S \cup B_S|}$$

Here A_S and B_S are the sets of sentences of the two summaries of the two summarization systems. The Jaccard coefficient is defined as the size of the intersection divided by the size of the union of the two sets. (Radev et al., 2002) We calculate the Jaccard coefficient for each of the 50 summaries created and then average the values to obtain an overall score.

6.2 Word Level comparison

To compare the degree of overlap at a finer level of granularity we investigate the overlap at word level as well. Here we computed the Jaccard coefficient (Nenkova et al., 2007) for sets of words generated by either of the summarization systems. Similar to the approach we followed for sentences, here too we calculate the Jaccard coefficient for each of the 50 topics and then average the score to get the overall Jaccard coefficient score at word level.

Table.3 shows the overall Jaccard coefficient scores for the MMR and LexRank summarization systems. The Jaccard coefficient for word level comparison shows very moderate similarity of 12.96%. The sentence level similarity was also very low at 1.68%.

7 Related Work

In this section we discuss a few acknowledged summarization paradigms that are based on the same approaches as our implementation in this paper. We also discuss the pros and cons of some of the State-of-the-art summarization techniques.

Boudin et al.s work named a scalable MMR approach to sentence scoring for multi-document update summarization is an extension of our re-implementation of Carbonell, J. and J. Goldstein 1998. In their work, Bounding et al. score the sentences by combining query relevance and dissimilarity with already read documents. As the amount of data in history increases, non-redundancy is pri-

	1	2	3	4	5	6	7	8	9	10	11
1	1	0.45	0.02	0.17	0.03	0.22	0.03	0.28	0.06	0.06	0
2	0.45	1	0.16	0.27	0.03	0.19	0.03	0.21	0.03	0.15	0
3	0.02	0.16	1	0.03	0	0.01	0.03	0.04	0	0.01	0
4	0.17	0.27	0.03	1	0.01	0.16	0.28	0.17	0	0.09	0.01
5	0.03	0.03	0	0.01	1	0.29	0.05	0.15	0.2	0.04	0.18
6	0.22	0.19	0.01	0.16	0.29	1	0.05	0.29	0.04	0.2	0.03
7	0.03	0.03	0.03	0.28	0.05	0.05	1	0.06	0	0	0.01
8	0.28	0.21	0.04	0.17	0.15	0.29	0.06	1	0.25	0.2	0.17
9	0.06	0.03	0	0	0.2	0.04	0	0.25	1	0.26	0.38
10	0.06	0.15	0.01	0.09	0.04	0.2	0	0.2	0.26	1	0.12
11	0	0	0	0.01	0.18	0.03	0.01	0.17	0.38	0.12	1

Table 2: Intra-sentence cosine similarities for documents in Table.1

	ROUGE-1			**ROUGE-2**			**ROUGE-SU4**		
System	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
CENTROID	35.94	36.41	36.16	7.86	7.97	7.91	12.23	12.40	12.31
DPP	39.70	39.79	39.74	9.59	9.62	9.60	13.82	13.85	13.84
ICSISUMM	38.58	38.41	38.48	9.84	9.78	9.81	13.69	13.62	13.65
LEXRANK	35.90	35.95	35.92	7.45	7.47	7.46	11.90	11.92	11.91
SUBMODULAR	39.25	39.18	39.21	9.36	9.35	9.35	13.77	13.74	13.75
MMR	32.33	32.5	32.41	5.22	5.25	5.24	9.75	9.81	9.78
OUR LEXRANK	29.29	28.99	29.12	4.63	4.58	4.6	9	8.91	8.95

Table 3: Summarization results evaluated by ROUGE (%).

Word level (%)	12.96
Sentence level (%)	1.05

Table 4: Jaccard coefficient - MMR vs LexRank

oritized over query-relevance and a promising performance on the DUC 2007 update corpus has been shown. There are many interesting extensions to Erkan and Radevs LexRank approach which has also been re-implemented in our work. Some of the most interesting ones have been discussed below:

Biased Random Walks (Erkan, 2006) by Erkan himself is a version of the LexRank algorithm implemented on DUC 2004 and extended to the focused summarization task of DUC 2006. As in LexRank, biased random walks represent the set of sentences in a document cluster as a graph, where nodes are sentences and links between the nodes are induced by a similarity relation between the sentences. Then ranks are given to the sentences according to a random walk model defined in terms of both the inter-sentence similarities and the similarities of the sentences to the topic description.

Li et al.s work on **aspect-oriented multi-document summarization** (Garg et al., 2009) is another paradigm that is based on our re-implementation of LexRank and introduces the event-aspect LDA model to cluster sentences into aspects. We then use extended LexRank algorithm to rank the sentences in each cluster. Integer Linear Programming for sentence selection has been used in their work with key features being automatic grouping of semantically related sentences and sentence ranking based on extension of random walk model. A new sentence compression algorithm which use dependency tree instead of parser tree was implemented as a part of their work and was compared with four baseline methods.

Yet another interesting extension to our approach is Garg et al.s work that presents an unsupervised,

graph based approach for extractive summarization of meetings. Graph based methods such as TextRank have been used for sentence extraction from news articles in their model with sentences as nodes and edges based on word overlap. A sentence node is then ranked according to its similarity with other nodes. Their algorithm clusters the meeting utterances and uses these clusters to construct the graph.

Some of the other novel sophisticated approaches for generic multi-document summarization have been discussed below:

CLASSY 04 [Peer 65]: This approach by Conroy et al. (Conroy et al., 2004) was the best among those that entered the official DUC 2004 evaluation. It is often used as comparison system by developers of novel summarization methods. It employs a Hidden Markov Model, using topic signature as the only feature. The probability of one sentence being selected in the summary also depends on the importance assigned to its adjacent sentences in the input document. It is worth noting that there is a linguistic pre-processing component in this system.

DPP: Determinantal point processes by Kulesza and Taskar (Kulesza et al., 2012) are probabilistic models of sets which balance the selection of important information and diverse groups of sentences within a given length. Specifically, DPPs combine a per-sentence quality model that prefers relevant sentences with a global diversity model encouraging non-overlapping content. This setup has several advantages. First, by treating these opposing objectives probabilistically, there is a rigorous framework for trading off between them. Second, the sentence quality model can depend on arbitrary features, and its parameters can be efficiently learned from reference summaries via maximum likelihood training; in contrast, most standard summarization techniques are tuned by hand. Finally, because a DPP is a probabilistic model, at test time it is possible to sample multiple summaries and apply minimum Bayes risk decoding, thus improving ROUGE scores. The DPP model in this work is trained on the DUC 2003 data to optimize the ROUGE-1 F-score.

RegSum: The RegSum system by Hong and Nenkova (Hong and Nenkova, 2014) employs a supervised model for predicting word importance. This model is superior to prior methods for identifying the words which are included in human models. RegSum combines the weights estimated from three unsupervised approaches, along with features including locations, part-of-speech, name-entity-tags, topic categories and contexts. Specifically, this system captures words which are of intrinsic interest to people by analysing a large number of summary-abstract pairs from the New York Times corpus by Sandhaus (Sandhaus, 2008). The summarizer employs the same greedy optimization framework as FreqSum and TsSum. It shows that the quality of the summaries could be greatly improved by better estimation of word importance.

Submodular: Treating multi-document summarization as a submodular maximization problem has proven successful by Lin and Bilmes (Lin and Bilmes, 2011) and has spurred a great deal of interest in this line of research by Sipos et al.(Sipos et al., 2012) and Morita et al., (Morita et al., 2013) as mentioned in the work by Hong et al. (Hong et al., 2014) in comparison for a repository of state of the art and competitive baseline summaries for generic news summarization. The advantage of using a submodular function to estimate summary importance is that there is an efficient algorithm for incrementally computing the importance of a summary with a performance guarantee on how close the approximate solution will be to the globally optimal one.

8 Conclusion

The rate of information growth due to the World Wide Web has called for a need to develop efficient and accurate summarization systems. Although research on summarization started about 50 years ago, there is still a long trail to walk in this field. Over time, attention has drifted from summarizing scientific articles to news articles, electronic mail messages, advertisements, and blogs. Our work provides an insight into glossed-over automatic text summarization approaches on their implementation simplicity and the speed of summary generation using statistical methods. Since a lot of interesting work is being done far from the mainstream research in this field, we have chosen to include a brief discussion on some acknowledged methods that we found interesting to allow us to carry out a unique comparison between existing approaches. We have

also outlined the methodology for reporting results, establishing informed choices for ROUGE settings and for the computation of statistical significance.

It is quite evident from our experiments that MMR ranking provides a useful and beneficial manner of providing information to the user by allowing the user to minimize redundancy. This is especially true in the case of query-relevant multi-document summarization providing a great scope for future work. We also re-implement the LexRank approach to define sentence salience based on graph-based centrality scoring of sentences with three different methods for computing centrality in similarity graphs. Even the simplest approach we have taken, degree centrality, is a good enough heuristic to perform better than lead-based and centroid-based summaries.

As an extension to our work, for a more sematic comparison between MMR and LexRank by transforming the sentences in the extracted summaries to sentence content units (SCUs) as in the pyramidal evaluation and then measuring the degree of similarity.

References

Florian Boudin and Juan-Manuel Torres-Moreno. 2007. A cosine maximization-minimization approach for user-oriented multi-document update summarization. In *Recent Advances in Natural Language Processing (RANLP)*, pages 81–87.

Ronald Brandow, Karl Mitze, and Lisa F Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing & Management*, 31(5):675–685.

John M Conroy, Judith D Schlesinger, Jade Goldstein, and Dianne P Oleary. 2004. Left-brain/right-brain multi-document summarization. In *Proceedings of the Document Understanding Conference (DUC 2004)*.

Dipanjan Das and André FT Martins. 2007. A survey on automatic text summarization. *Literature Survey for the Language and Statistics II course at CMU*, 4:192–195.

Brigitte Endres-Niggemeyer, Ricklinger Stadtweg, and Brigitte Endres. 2000. Human-style www summarization. *Report. Hannover: University of Applied Sciences and Arts*.

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

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

Günes Erkan. 2006. Using biased random walks for focused summarization. *Ann Arbor*, 1001:48109–2121.

Nikhil Garg, Benoit Favre, Korbinian Reidhammer, and Dilek Hakkani Tür. 2009. Clusterrank: a graph based method for meeting summarization. Technical report, Idiap.

Albert Gatt and Ehud Reiter. 2009. Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 90–93. Association for Computational Linguistics.

Pierre-Etienne Genest, Guy Lapalme, and Mehdi Yousfi-Monod. 2013. Hextac: the creation of a manual extractive run. *Génération de résumés par abstraction*, page 7.

Kai Hong and Ani Nenkova. 2014. Improving the estimation of word importance for news multi-document summarization. In *EACL*, pages 712–721.

Kai Hong, John M Conroy, Benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A repository of state of the art and competitive baseline summaries for generic news summarization. In *LREC*, pages 1608–1616.

Hongyan Jing and Kathleen R McKeown. 2000. Cut and paste based text summarization. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 178–185. Association for Computational Linguistics.

Alex Kulesza, Ben Taskar, et al. 2012. Foundations and trends® in machine learning.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 510–520. Association for Computational Linguistics.

Ziheng Lin, Tat-Seng Chua, Min-Yen Kan, Wee Sun Lee, Long Qiu, and Shiren Ye. 2007. Nus at duc 2007: Using evolutionary models of text. In *Proceedings of Document Understanding Conference (DUC)*.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Rada Mihalcea, Paul Tarau, and Elizabeth Figa. 2004. Pagerank on semantic networks, with application to word sense disambiguation. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1126. Association for Computational Linguistics.

Hajime Morita, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2013. Subtree extractive summarization via submodular maximization. In *ACL (1)*, pages 1023–1032. Citeseer.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2):4.

Dragomir R Radev, Eduard Hovy, and Kathleen McKeown. 2002. Introduction to the special issue on summarization. *Computational linguistics*, 28(4):399–408.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Ruben Sipos, Pannaga Shivaswamy, and Thorsten Joachims. 2012. Large-margin learning of submodular summarization models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 224–233. Association for Computational Linguistics.

René Witte, Ralf Krestel, and Sabine Bergler. 2007. Generating update summaries for duc 2007. In *Proceedings of the Document Understanding Conference*, volume 2007.