CORPORA EVALUATION AND SYSTEM BIAS DETECTION IN MULTI DOCUMENT SUMMARIZATION

Student Name: Alvin Dey
Roll Number: IIIT-D-MTech-CS-20-MT18066
June, 2020

Indraprastha Institute of Information Technology
New Delhi

Thesis Advisors
Dr. Tanmoy Chakroborty

Submitted in partial fulfillment of the requirements
for the Degree of M.Tech. in Computer Science & Engineering

©2020 IIIT-D-MTech-CS-20-MT18066
All rights reserved

Certificate

This is to certify that the thesis titled **"Corpora Evaluation and System Bias Detection in Multi-document Summarization"** submitted by **Alvin** for the partial fulfillment of the requirements for the degree of Master of Technology in Computer Science & Engineering is a record of the bonafide work carried out by him under my guidance and supervision at Indraprastha Institute of Information Technology, Delhi. This work has not been submitted anywhere else for the reward of any other degree.

.....

Dr. Tanmoy Chakraborty

Indraprastha Institute of Information Technology, New Delhi

Abstract

Multi-document summarization (MDS) is the task of reflecting key points from any set of documents into a concise text paragraph. In the past, it has been used to aggregate news, tweets, product reviews, etc. from various sources. Owing to *no standard definition of the task*, we encounter a plethora of datasets with varying levels of overlap and conflict between participating documents. There is also *no standard regarding what constitutes summary information* in MDS. Adding to the challenge is the fact that new systems report results on a set of chosen datasets, which might not correlate with their performance on the other datasets. In this paper, we study this heterogeneous task with the help of a few widely used MDS corpora and a suite of state-of-the-art models. We make an attempt to quantify the quality of summarization corpus and prescribe a list of points to consider while proposing a new MDS corpus. Next, we analyze the reason behind the absence of an MDS system which achieves superior performance across *all* corpora. We then observe the extent to which system metrics are influenced, and bias is propagated due to corpus properties.

Acknowledgment

I would like to express my utmost gratitude towards my advisor Dr. Tanmoy Chakroborty for his guidance and support. The quality of this work is due to his immensely helpful advice. I thank Tanya Chowdhury and Yash Kumar Atri for their support and collaboration in this work. Last but not the least, I would like to thank my supportive family who encouraged and kept me motivated throughout the project.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Any technique to compress more than one document into a textual summary is referred to as multi-document summarization (MDS).

It has a lot of applications in day-to-day life – gaining insight from tweets related to a hashtag, understanding product features amongst e-commerce reviews, summarizing live blogs related to an ongoing match, expressing popular opinion amongst the users of online service forums, etc.

Most study on MDS was performed during the DUC and TAC challenges organized by NIST starting in the early 2000s. Each version of the challenge had its own dataset, and most systems submitted to these challenges were unsupervised and extractive. Over time the DUC and TAC challenge data became the most popular corpora for multi-document summarization. These datasets were manually accumulated and had less than a hundred instances each.

Since the deep learning boom, there has been a significant increase in the number of supervised document summarization systems. Large labeled corpora which are not manual but crowd-sourced have been introduced to meet the training requirements of these supervised systems. However, these crowdsourced corpora widely differ in quality based on factors like genre, size of the community, presence of moderation in the community, etc.

In single-document summarization (SDS), supervised neural systems outperform unsupervised non-neural extractive approaches by a significant margin (measured by the ROUGE score). Popular crowdsourced datasets such as CNN/Dailymail, Gigaword and NYT are large enough to train supervised algorithms and have become the de facto standard for SDS. However, there

is a lack of such a standard in MDS. This may be attributed to the complexity of the problem in comparison to SDS, the hardness of accumulating labeled data or more so in the definition of what constitutes a multi-document summary.

In recent times, a few large scale datasets for MDS have been introduced. However, there has been no study to measure the relative complexity of these datasets. With most of them being crowdsourced, it becomes an important problem to gauge their quality. In this paper, *we make a pioneering attempt to quantify the quality of an MDS dataset.* We provide a list of corpus-based metrics that should be necessarily reported by the proposers of new MDS corpora in order to enable comparison with existing datasets.

Although summarization systems today have been improving in terms of ROUGE as compared to systems a decade ago, manual inspection reveals that they tend to portray more and more biases. Usually, new systems are introduced with results on a single corpora, making it difficult to understand if the biases are introduced by the system or are inflected by the corpora used for training. We find common biases shown by new age neural models and examine the extent to which training data affects such biases exhibited by the summarization systems.

We study MDS corpora – DUC [6], TAC [25], Opinosis [9], Multinews [8], CQASumm [4], and popular summarization systems – LexRank [7], TextRank [20], MMR [3], ICSISumm [12], PG [24], PG-MMR [18], Hi-MAP [8], and Transformer [11].

We define multi-document summarization as a mapping from a set of non-independent candidate documents to a synopsis which covers *important* and *redundant* content present in the source. We propose metrics to model the quality of an MDS corpora in terms of - Abstractness, Inter Document Similarity (IDS), Redundancy, Pyramid Score, Position Bias and Multi-Document Summarization Coverage (MDS Cov). We also present metrics to capture biases displayed by summarization systems and then look for strong correlations between them and their respective corpora metrics. We present a comprehensive framework for new corpora proposers, where we highlight that summarization corpus quality is a function of two distinct features – the quality of candidate documents and the quality of reference summaries. We further mathematically model these two features and give an end-to-end metric for authors of new corpora to report. To the best of our knowledge, we are the first work in this direction.

1.2 Research Questions

Q1. How should one model the quality of a MDS corpus as a function of its intrinsic properties?
Q2. Why do the ROUGE-based ranks of different MDS systems differ across different corpora? How should an MDS system which indents to achieve high ROUGE scores across *all* corpora, look like?
Q3. Why do systems show bias on different metrics, and which other system and corpus attributes are the reason behind it?
Q4. Is the task of MDS almost solved, or there is still plenty of scopes for improvement?

1.3 Terminology Used

• **Candidate Documents**: Set of document under a given topic of a multi-document summarization corpus.
• **Reference Summary**: Ground truth summary provided for each topic in a given corpus.
• **System Summary**: Summary generated by a summarization system for a given topic in a corpus.
• **Semantic Units**: An atomic piece of information.
• **Coherence Score**: An intrinsic metric used in topic models to justify hyperparameter choice.

1.4 Thesis Outline

The rest of the thesis is organized as follows:

Chapter 2 : Related works in the field
Chapter 3 : Datasets
Chapter 4 : Systems
Chapter 5 : Corpus and System metrics proposed
Chapter 6 : Experiments conducted and Results
Chapter 7 : Inferences and Conclusion
Chapter 8 : Future work

Chapter 2

Related Works

Previous attempts to evaluate the quality of a summarization corpora are few in number and mostly from the time when corpora were manually accumulated. [14] primarily use the intrinsic metrics of precision and recall to evaluate corpora quality. In addition, they use an extrinsic metric, called Pseudo Question Answering. This metric evaluates whether a summary has an answer to a question that is otherwise answerable by reading the documents or not. Although effective, the cost of such an evaluation is enormous and is not scalable to modern day corpora sizes. In corpora where multiple references are available. They highlight important sentences within each document of a corpus and also those documents with the same content is annotated as an additional feature of the corpus. They evaluate redundancy as a corpus metric and discuss how it affects the corpus quality and reducing it might aid in the efforts of summarization performance improvement. [1] use inter-annotator agreement to model the quality of the corpora. They also use non-redundancy, focus, structure, referential clarity, readability, coherence, length, grammaticality, spelling, layout, and overall quality as quantitative features for a multi-document summarization corpora. They use non-uniform corpus to focus on issue that most knowledge seekers usually face – the problem of distinctive sources. They extract information for multiple different source but organize and redact it minimally to form a coherent set of text. Their work highlights the importance of coherence while judging corpus quality, as quantitative metrics undermines the importance of corpus quality in terms of meaningfulness.

There have been a series of very recent works that look into how to strengthen the definition and discover system biases in single-document summarization. Neural summarization and it's recent

advancements in the scope of single-document summarization on the frameworks dependency on the corpus. Based on this very idea , [16] study how position, diversity and importance are significant metrics in analysing the toughness of nine popular single-document summarization corpora. Their findings exhibit significant position bias in corpora built on news article, while no such sign of similar trend is observed in corpora built on meeting minutes and academic papers. Their empirical methods shows that different neural based summarization systems shows accumulation of position, diversity and importance in various degrees. Their study highlights important aspects for consideration while proposing a new summarization corpora or building a new system. Another recent work [17] extensively study the position bias in news datasets that most single-document summarization system seem to exploit. They evaluate the underlying reasons of stagnancy in performance summarization benchmark datasets. They assess the dataset, evaluation metrics and models. They highlight that shortcoming are 3-folds – datasets which have been crowdsourced or scrapped from various websites contain noise of varying degrees and need to be cleaned in a more proportionate way to be used as benchmark, current evaluation metrics for summarization doesn't correlate with human opinion on the same as well as it should, neural architecture tend to model biases based on their layout which hinders their understanding of what an actual summary should be.

Two seminal works – [22] and [23] exploit the theoretical complexity of summarization on grounds of importance, analyzing in depth what makes for a good summary. [22] focuses on answering the more fundamental questions on summarization. It narrows summarization down to 3 aspects – Redundancy, Relevance and Informativeness. They formalize importance as a function of the above 3. They provide interpretations of these metrics and experiments to guide future research in the field. [23] observes that automatic evaluation metrics of summaries such as ROUGE are well received because they are perceived to correlate with human judgments. They provide intuition as to why this might not be the case with modern systems, as they are convincingly better than the best systems submitted during the DUC/TAC shared tasks era. Their work quantitatively shows how human judgements only correlate with automatic metrics in the average scoring range, they deviate from human judgements in the high scoring area where modern summarization systems operate. This research raises questions on present evaluation metrics of summarization and probes for research on them. [13] propose a new single document summarization corpora of over 1.2 million news articles an summaries written by human annotators of newsroom of 38 different news publications. The summaries provided as ground truth combines

the extractive and abstractive essence of summarization. They quantify how it compares to other datasets in terms of diversity and difficulty of the data. They use extractive fragment coverage, extractive fragment density and compression ratio to analyse the datasets.

Systems such as the ones proposed by [27], uses position bias as a method for training their model. They score better than many competitive baseline models on news articles datasets. To the best of my knowledge, no comparative work exists for either corpora or systems in MDS.

Chapter 3

Datasets

In this work, we study some widely used multi-document summarization corpora from different genres.

3.1 DUC

DUC [6] is a news dataset built using newswire/paper documents. The 2003 and 2004 versions have summaries (<= 665 bytes), written by NIST assessors for each topic cluster. NIST staff chose 30 and 50 topics for 2003 and 2004 versions respectively, for which TDT annotators collect an average 10 documents under each topic in both versions. They have 30 and 50 topics respectively with each topic having 4 manually curated reference summaries.

3.2 TAC

TAC [25] is built from the AQUIANT-2 collection of newswire articles where NIST assessors select 48 and 44 topics for the 2008 and 2010 versions, respectively. Each topic contains an average of 10 articles, with four manually sourced references each. 20 articles were selected for each topic, which were divided under initial and update task with 10 articles under each topic. TAC summarization dataset is a continuation of the DUC summarization tasks on news articles. We use the 2008 and 2010 versions of the dataset for our work. The summarization task is divided in two segments: Initial and Update. Initial consists of 100 words summary for each topic. Update consists of a 100 words summary for subsequent documents in the same topic, with the assumption that reader has read the documents in the Initial task. Both the tasks

consist an average of 10 articles under each topic.

3.3 Opinosis

Opinosis [9] is an accumulation of user reviews collected from various sources like TripAdvisor, Edmunds.com and Amazon. There are 51 topics, with each topic having approximately 100 sentences on an average. There are about 4 human written summaries for each topic.

3.4 CQASumm

CQASumm [4] is a community questions answering dataset, curated by filtering 4.4 million threads from the Yahoo! Answers L6 dataset. It treats each answer under a thread as a separate document and modifies the best answer when applicable to generate a reference summary. It ignores threads with less than 5 answers, best answers with less than 100 words and answers other than best answer with less than 200 words. The reference summaries with more than 100 words are restricted to 100 words by sorting sentences with priority based on it's content. For threads with popular opinion in answers, which is not not reflected by the best answer, they use a graph based cumulative similarity approach to filter them out. For threads containing opinion not present in other answers, they use a correlation criteria to drop those threads. The corpus has nearly 100,000 question threads with a total of 1,201,744 answers. Threads have an average of 12 answers, with each approximately 65.03 words and a 100 word upper limit for summary.

3.5 Multinews

Multinews [8] is a news dataset comprised of news articles and human-written summaries from newser.com. Each summary is written by professional editors. It consists news from over 1500 sites, containing a total of 2,50,000 articles which was downloaded and processed to suit MDS. Final dataset has 56,216 topics, with summaries of 260 words on average. A total of 20 editors contribute to 85% of the summaries.

Cambodian leader Hun Sen on Friday rejected opposition parties' demands for talks outside the country, accusing them of trying to "internationalize" the political crisis. Government and opposition parties have asked King Norodom Sihanouk to host a summit meeting after a series of post-election negotiations between the two opposition groups and Hun Sen's party to form a new government failed.
King Norodom Sihanouk has declined requests to chair a summit of Cambodia's top political leaders, saying the meeting would not bring any progress in deadlocked negotiations to form a government. Cambodian leader Hun Sen's ruling party and the two-party opposition had called on the monarch to lead top-level talks, but disagreed on its location.
Cambodia's two-party opposition asked the Asian Development Bank Monday to stop providing loans to the incumbent government, which it calls illegal.
Cambodia's ruling party responded Tuesday to criticisms of its leader in the U.S. Congress with a lengthy defense of strongman Hun Sen's human rights record. The Cambodian People's Party criticized a non-binding resolution passed earlier this month by the U.S.
Cambodia's leading opposition party ruled out sharing the presidency of Parliament with its arch foe Saturday, insisting it alone must occupy the top position in the legislative body.
Prospects were dim for resolution of the political crisis in Cambodia in October 1998. Prime Minister Hun Sen insisted that talks take place in Cambodia while opposition leaders Ranariddh and Sam Rainsy, fearing arrest at home, wanted them abroad. King Sihanouk declined to chair talks in either place. A U.S. House resolution criticized Hun Sen's regime while the opposition tried to cut off his access to loans. But in November the King announced a coalition government with Hun Sen heading the executive and Ranariddh leading the parliament. Left out, Sam Rainsy sought the King's assurance of Hun Sen's promise of safety and freedom for all politicians.

Table 3.1: A sample DUC Thread

The A380, the new Airbus "superjumbo" which will be officially unveiled Tuesday, is the product of a decade of designing, drumming up of advance orders and dreams of knocking Boeing's 747 off its perch as the top bird in passenger transport.
The A380 'superjumbo', which will be presented to the world in a lavish ceremony in southern France on Tuesday, will be profitable from 2008, its maker Airbus told the French financial newspaper La Tribune.
Airbus unveiled the world's biggest passenger jet Tuesday at a spectacular sound-and-light ceremony in Toulouse, southern France, where the leaders of France, Britain, Germany and Spain cheered Airbus's latest victory over Boeing for aviation dominance.
European airplane maker Airbus "is likely to discuss before the end of the year" a possible increase in production capacity of its new super-jumbo A380 aircraft, Airbus' production chief Gustav Humbert said in a magazine interview released Tuesday.
US complaints about European government subsidies to Airbus Industrie have not stopped US airports preparing for the arrival of the giant A380 jet that is set to spark a new transatlantic trade dogfight.
The European Airbus A380 flew its maiden test flight from France 10 years after design development started.The A380 super-jumbo passenger jet surpasses the Boeing 747 and breaks their monopoly. Airlines worldwide have placed orders but airports may need modification to accommodate the weight and width of the A380. U.S. airlines have not placed an order. Airbus has fallen behind in production and a backlog of orders has developed. Airbus must sell at least 250 planes to break even financially. The A380 is overweight and modifications to meet the weight requirements impacted the budget. Additional test flights are planned.

Table 3.2: A sample TAC Thread

Question : Who will be the World Cup Champion ?
Dude, American's don't follow "football" they way they do in the rest of the world. We're even goofy enough to change the name of the sport to soccer!
germany germany hey i come from india. but u know germany's the best they can bet argentina and brazil easily.
Best teams I've seen: Argentina & Spain, but cannot write off Germany - home fans, refreshingly exciting. And of course Brazil are the best team, so will eventually win
I don't think Brazil is gonna win this year i mean yeah i know they have good players but did you see their match against Croatia i mean alot of people were disappointed well not me cause i don't like them but anyway i think that its gonna be Germany who wins this year or maybe the Czech Republic
ARGENTINA, no doubt. They really play as a team, without selfishness. They have amazingly skillful players, even on the bench. The members of the team are very close. Pekerman is an excellent coach. And most importantly, they are ready to leave their heart on the field.
England : they were just lucky so far , nothing grand to be honest , and on the other hand they've never won any World Cup , and it does not look like it will change. - Checz Republic : They have been playing great " closet " football (I mean without the rest of the world being aware of it or without making a big fuzz) , so they might as well just be the great surprise. - Spain and Italy : They are big football names , and Spain played really well against Ukraine... OH

Table 3.3: A sample CQASumm Thread

The Da Vinci Code has sold so many copies—that would be at least 80 million—that it's bound to turn up in book donation piles. But at one charity shop in the UK, it's been donated so heavily that the shop has posted a sign propped up on a tower of Da Vinci Code copies that reads: "You could give us another Da Vinci Code... but we would rather have your vinyl!" The manager of the Oxfam shop in Swansea tells the Telegraph that people are laughing and taking pictures of the sizable display: "I would say that we get one copy of the book every day."
A woman reads a copy of the newly released book "The Lost Symbol" by Dan Brown, at a speed reading book launch event in Sydney, September 15, 2009. REUTERS/Tim Wimborne. SAN FRANCISCO The latest novel from "Da Vinci Code" author Dan Brown, "The Lost Symbol," broke one-day sales records, its publisher and booksellers said. Readers snapped up over one million hardcover copies across the United States, Canada and the United Kingdom after it was released on Tuesday, said publisher Knopf Doubleday, a division of Random House Inc. "We are seeing historic, record-breaking sales across all types of our accounts in North America for 'The Lost Symbol," said Sonny Mehta, editor in chief of Knopf Doubleday Publishing Group.
A charity shop is urging people to stop donating The Da Vinci Code after becoming overwhelmed with copies. The Oxfam shop in Swansea has been receiving an average of one copy of the Dan Brown novel a week for months, leaving them with little room for any other books.
Bestselling author is also the most frequently given away to charity shops. Dan Brown might be one of the world's bestselling authors but it turns out that readers aren't too keen on keeping his special blend of religious conspiracy and scholarly derring-do on their shelves once they've bought it.
Whether a sign of a good read; or a comment on the 'pulp' nature of some genres of fiction, the Oxfam second-hand book charts have remained in The Da Vinci Code author's favour for the past four years. Dan Brown has topped Oxfam's 'most donated' list again, his fourth consecutive year.
The Da Vinci Code was published in 2003, and within six years Brown had booted John Grisham from the No. 1 slot on the list of writers whose books were most often donated to Oxfam's 700 shops, reported the Guardian at the time. The Independent in 2012 reported Brown's best-seller was the most-donated book for the fourth year running. (See why Dan Brown took heat from the Philippines.)

Table 3.4: A sample Multinews Thread

Chapter 4

Systems

To identify bias in system generated summaries, we study a few non-neural extractive and neural abstractive summarization systems.

4.1 Extractive Systems

4.1.1 LexRank

[7] is a graph based algorithm that computes the importance of a sentence using the concept of eigen vector centrality in a graphical representation of text. All sentences are represented as graph nodes and edges represent the similarity between the two nodes. The similarity is calculated using frequency of the word in the sentence which is formally defined as

$$sim(x,y) = \frac{\sum_{w \in x,y} tf_{w,x} tf_{w,y} (idf_w)^2)}{\sqrt{\sum_{x_i \in x} (tf_{x_i,x} idf_{x_i})^2} \times \sqrt{\sum_{y_i \in y} (tf_{y_i,y} idf_{y_i})^2}} \tag{4.1}$$

4.1.2 TextRank

[20] represents the sentences as a fully connected graph where nodes are the sentences and edges represents a similarity score computed by a derived version of PageRank [2]. how similar they are. TextRank assumes all weights between the sentences to be unit weights and then uses a derived version of PageRank [2] to rank sentences.

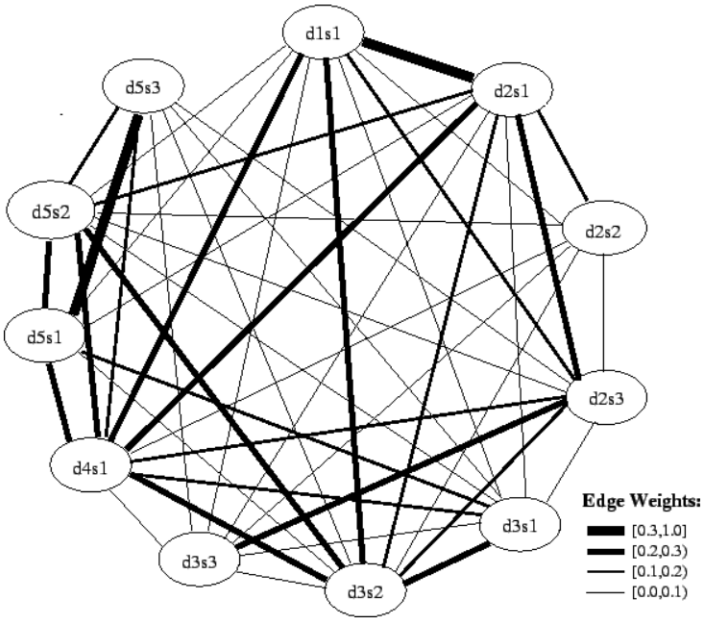


Figure 4.1: Graph highlighting edge weights and Node importance in LexRank

4.1.3 Maximal Marginal Relevance (MMR)

[3] is a popular extractive summarization system which select sentences based on higher relevance while maintaining redundancy. It reduces redundancy while maintaining relevant sentences in ranking sentences. MMR calculates the relevance of a query to sentences using a similarity metric. Similarity within sentences is calculated and similar sentences are removed.

4.1.4 ICSISumm

[12] optimizes the summary coverage by adopting linear optimization framework. It finds a globally optimal summary using the most important concepts covered in the document. It extracts sentences that contain frequent bigrams from the corpus. Given an upper bound on summary length, it optimizes the summary using integer linear programming. It finds a globally optimal summary using the most important concepts covered in the document.

4.2 Abstractive Systems

4.2.1 Pointer Generator (PG)

[24] network is one of the most popular sequence to sequence based summarization architectures. The pointer Generator architecture allows both copying words from the source text by pointing

or generating words from a fixed vocabulary. In PG network, encoder reads source document tokens w_i which produces sequence of encoder hidden states h_i, at each timestep t_i in decoder state s_t; decoder gets the previous word embeddings. Attention distribution a_t is calculated as follows:

$$e^{t}_{i} = v^{T}(W_{h}h_{i} + W_{s}s_{t} + b_{attn})$$
$$a^{t} = softmax(e^{t}), h^{*}_{t} = \sum_{i} a^{t}_{i}h^{t}_{i}$$

Here v, W_h, W_s and b_{attn} are all learnable parameters. The p_{gen} acts as a switch between generating words from the vocabulary or copying words from the source. The vocabulary distribution and copy probability are weighted, and the final distribution is obtained for summary generation. Context vector h^{*}_{t} is calculated using the attention distribution. A generation probability $P_{gen} = [0,1]$ is calculated which gives the probability of either generating words from the vocab or copy words from the source file. The vocabulary distribution and attention distribution are weighted and final distribution is obtained for summary generation. The encoder reads the text to generate a sequence of encoder hidden states. The decoder at each step gets the previous token that goes in the generated summary, this is used to update the hidden states which gives the attention distribution over words in the source text which is further used to calculate context vector and finally a pointer generator to copy words from source text if and when required.

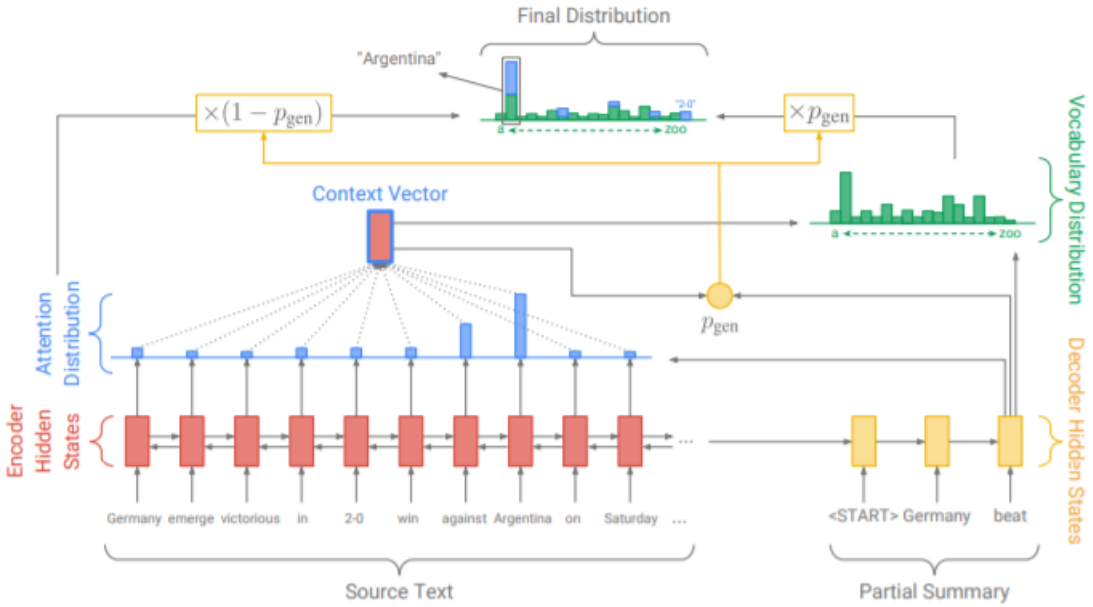


Figure 4.2: Pointer-Generator Architecture

4.2.2 Pointer Generator-MMR:

PG-MMR [18] uses MMR along with PG for better coverage and redundancy mitigation. Here MMR computes a similarity score of sentences with source text and modifies the attention weights for better summary generation. At each iteration in MMR, it selects one sentence s_i from the source document by finding the similarity with the source candidate set as follows:

$$\underset{s_i \in D \setminus S}{\operatorname{argmax}} [\lambda Sim_1(s_i, D) - (1 - \lambda) \underset{s_j \in S}{\max} Sim_2(s_i, s_j)]$$

where $Sim_1(s_i, D)$ measures the similarity between the sentence s_i with Document D. $max_{sj} \in S$ $Sim_2(s_i, s_j)$ approximates the maximum similarity between s_i with each of the summary sentences s_j.

Here Sim_2 acts as a proxy of redundancy and λ acts as a balancing factor.

At each iteration in PG-MMR, K source sentences are selected and the attention weights ($\alpha_{t,i}$) are dynamically adjusted at test time. It uses the same PG architecture but with MMR, Here MMR's work is to find K highest scored source sentence, The weights of all sentences are calculated and based on the weights the sentences that relate to the partial summary generated by the model receives low scores which helps the model to find the important content that might not have been included in the summary.

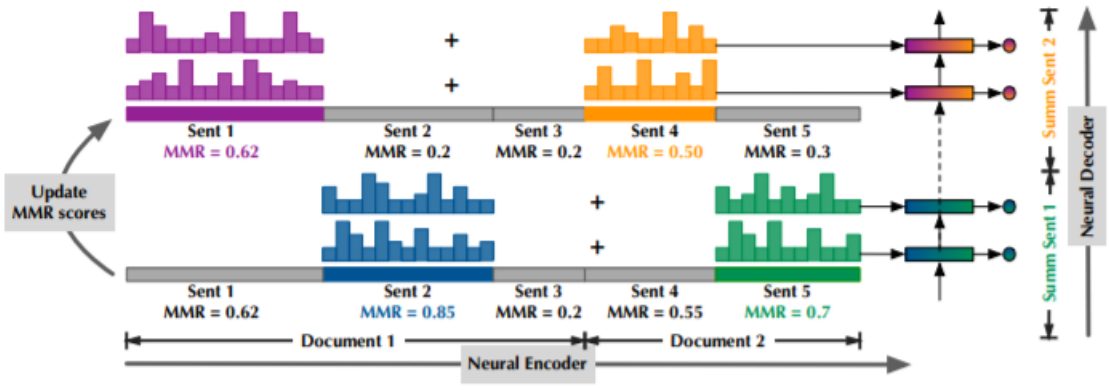


Figure 4.3: Pointer generator Maximal Marginal Relevance architecture

4.2.3 Hi-MAP:

Hierarchical MMR-Attention Pointer Generator model [8] extends the work of PG and MMR. Here MMR scores are calculated at word level and incorporated in the attention weights for better summary generation. For each sentence a ranking is computed using Equation 4.2.2. s_i, en-

coder output of last token is taken. Sentence embeddings are represented by $h_s^D = [h_1^S, h_2^s, ..., h_n^s]$. MMR rankings on these sentences h_s^D is computed by Equation 4.2.2 .

$$v_{ij} = tanh(h_j^{sT} W_{self} h_i^s)$$
$$B_{ij} = \frac{exp(v_i j)}{\sum_j exp(v_i j)}, score_i = \max_j(B_{i,j})$$

where s_{sum} represents the query vector, $W_{i,j}$ is a learnable parameter. Further MMR_i computes the sentence representation to determine the optimal sentence based on relevance and redundancy. For each sentence s_i, the normalized MMR_i scores are given by

$$\overline{MMR_i} = \frac{exp(MMR_i)}{\sum_i MMR_i}, \overline{a^t} = a^t \overline{MMR_i}$$

Attention weights for each token are given by $\overline{a_t}$ and final summaries are generated. Word tokens for the whole document are taken as a single input by the encoder and the encoder output for the last token is saved to obtain the representation of both source articles and summaries. MMR scores are computed for the sentence representations to determine the optimal sentences based on relevance and redundancy.

It is build on top of pointer generator model, Here it uses the same bi-lstm network used in PG and obtain the sentence level representation of both articles and summaries. The MMR computes the rankings on the candidate sentences and finds the salient sentences and returns a score which is further normalized by applying softmax function and the attention weights are updated.

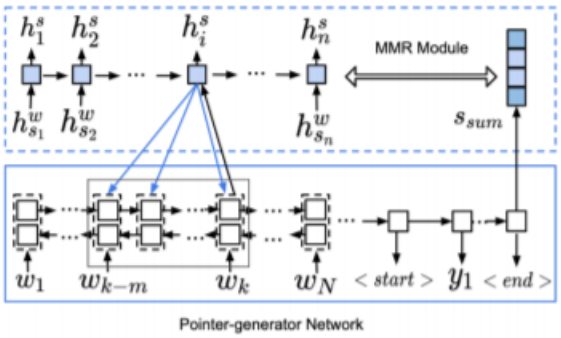


Figure 4.4: Hi-Map architecture

4.2.4 Bottom-up Abstractive Summarization (CopyTransformer)

[11] It uses the big transformer parameters proposed by [26] but here one of the attention heads chosen randomly acts as a copy distribution. It uses the encoder decoder structure but with stacked self-attention and point-wise fully connected layers. In encoder, each layer has two sub layers. The first is the multi-head self-attention mechanism, and the second is the fully connected layer. A residual connection between each of the sub layers is introduced followed by layer normalization. Decoder also houses two sub layers along with a third sub layer which performs multi-headed attention over the output of the encoder. In this architecture, one of the attention heads chosen randomly acts as a copy-distribution. It uses a transformer architecture with 512 dimensions for encoder and decoder. In this architecture one of the attention heads chosen randomly acts as a copy-distribution. This is an extension of Attention is all you need for abstractive summarization. It uses the original transfomer implementation with "In this architecture one of the attention heads chosen randomly acts as a copy-distribution." with this change only.

Chapter 5

Corpus and System metrics

5.1 Corpus Metrics

• **Abstractness**: This metric provides an insight about how many *new* tokens are present in the ground-truth summaries. We define *Abstractness* as the percentage of different higher order n-grams between the reference summary and candidate documents as follows:

$$Abstractness(s) = 1 - \frac{\sum_{i \in 1..m} C_n(S \cap D_i)}{C_n(S)} \tag{5.1}$$

where S represents the set of n-grams in the reference summary, D_i represents the set of n-grams in the i^{th} candidate document, m is the number of documents, and C_n represents the count of n-grams in a set.

• **Inter Document Similarity (IDS)**: This similarity score quantifies how identical two documents under a same topic are. We define it as the degree of topical overlap between documents under a topical cluster. We use the theoretical model of relevance [22] to calculate the similarity between topic distributions of two documents as follows:

$$IDS(D_i) = \frac{\sum_{D_j \in S} Relevance(D_j, D_i)}{|S|} \tag{5.2}$$

where D_i is the i^{th} document in a topic, S is the set of all documents other than D_i under the same topic. $Relevance(.,.)$ is defined as

$$Relevance(A, B) = \sum_{\omega_i} P_A(\omega_i).\log(P_B(\omega_i)) \tag{5.3}$$

where P_A represents the probability distribution of documents A, and ω_i represents the i^{th} semantic unit in the distribution.

• **Pyramid Score**: [21] define *Pyramid Score* as an approach to evaluate system summaries from the observed distribution of content in the pool of ground-truth summaries. We refine this metric to quantitatively analyse the reference summary provided in the dataset w.r.t. the important aspects of the documents. [21] define *Summarization Content Unit* (SCU) as a unit of information which is close to a clause in length and focuses on the crucial aspect inside a sentence. We divide the documents under a topic into a group of SCUs with a weight (e.g., number of unique documents containing that SCU) assigned to each SCU. We build the Pyramid of SCUs with the highest weight at the top and lowest at the bottom. For any summary S which contains a total of X SCUs to be evaluated, we first calculate the optimal score of any summary with X SCUs, given the pyramid constructed before as follows:

$$Optimal\,score = \sum_{i=j+1}^{n} i|T_i| + j(X - \sum_{i=j+1}^{n} |T_i|)$$

where n is the number of tiers in the pyramid, $|T_i|$ is the number of SCUs in tier T_i, and j denotes the lowest tier in the pyramid an optimal summary will be draw from, i.e., $j = \max_i(\sum_{t=i}^{n} |T_i| \geq X)$. The SCU weight of our S is: $Summary\,score = \sum_{i=1}^{n} i \times S_i$, where S_i is the number of SCUs that appear in tier T_i. Finally, the Pyramid score is defined as:

$$Pyramid\,score = \frac{Summary\,score}{Optimal\,score}$$

• **Redundancy**: Similar to [22], we define redundancy as the topical distribution skewness for candidate documents. It measures the amount of information present in the document in terms of its topical distribution.

$$Redundancy(D) = \sum_{\omega_i} P_D(\omega_i).\log(P_D(\omega_i)) \tag{5.4}$$

where P_D represents the probability distribution of documents D, and ω_i represents the i^{th} topic in the distribution.

• **Position Bias**: We define positional importance bias across a document as the degree of change in importance w.r.t. the ground-truth over the course of candidate documents. The idea relies on dividing the document into 'k' segments, calculating the importance of each segment w.r.t. the ground-truth by a similarity score.

$$PositionalImportance(D_j) = \max_{1 \leq i \leq n} sim(D_j, R_i) \tag{5.5}$$

where D_j is the j^{th} sentence in the document, R_i represents the i^{th} sentence in the reference. sim is a similarity metric between two sentences and n is the total number of sentences in the reference summary.

• **Multi-document Summarization Coverage (MDS Cov)**: We propose MDS Cov score to quantify the bias that a reference summary exhibits w.r.t. its set of candidate documents. It measures the importance given to each document in the candidate set by the reference summary as:

$$MDSCov(D, S) = Var_j\big(D_j \cap S_u\big) \tag{5.6}$$

Here, D and S are the set of candidate documents for MDS and their summary respectively, Var is the variance, D_j and S_u are the sets of SCUs in the j^{th} document of the candidate set and the reference summary respectively.

5.2 System Metrics

• **ROUGE**: [19] It is one of the widely used metrics for evaluating summaries. It measures the n-gram overlap between the system generated summary and the reference summary. ROUGE-N, which is an n-gram recall is calculated as: It measures the quality of a summary by comparing it with the reference summary. There are different sub metrics used in ROUGE like ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-S. ROUGE-N, which is an n-gram recall between the system generated summary and the reference summary, which is calculated as:

$$R_n = \frac{\sum_r \sum_s match(gram_{s,c})}{\sum_r \sum_s count(gram_s)} R_n = \sum_r \sum_s match(gram_{s,c}) \Big/ \sum_r \sum_s count(gram_s) \tag{5.7}$$

Here $\sum_s$ calculates how many times a n-gram appears in the candidate summary. In the presence of multiple reference summaries, $\sum_r$ iterates over all the available summaries. The denominator represents the count on n-grams in the reference summary.

• **F1 Score with Oracle Summaries**: Oracle summaries reflects the highest ROUGE score possible over the source document given the reference summary. The basic idea behind this metric is to study the efficiency of the summarization system i.e., whether the summaries generated contain more or less information than what is required.

• **System Abstractness**: The core insight behind abstractive summarization system is to create summaries by generating novel sentences either by rephrasing or introducing new words. Novel n-gram generation is calculated using Eq. 5.1. The n-grams of reference summary is replaced by system summaries in the above equation.

$$Coverage(D, S) = \frac{\sum_{i \in 1..n}(D \cap S_i)}{C_n(S)} \tag{5.8}$$

where D represents the set of n-grams in the article source document, and S represents the set of n-grams in the i^{th} system summary. The denominator denotes the total count of n-grams in a

system summary. Finally, the values of all articles is normalized to get the score for the system.

• **Position Distribution**: We study the behaviour of systems with reference to the presence of information in the source text. For this, we segment the document in certain parts uniformly and then compute the similarity of n-gram tokens of system summaries w.r.t. the source document segment.

• **Inter Document Similarity (IDS)**: For a given pair of documents, IDS quantifies how similar they are under a given topic. It computes a similarity score for each document against the topic. The relevance for system summaries is calculated by Eq. (5.3),

$$Relevance(A, B) = \sum_{\omega_i} P_A(\omega_i).\log(P_B(\omega_i)) \tag{5.9}$$

where P_A represents the probability distribution of system summary S, and ω_i represents the i^{th} semantic unit in the distribution.

$$IDS(D_i) = \frac{\sum_{D_j \in S} Relevance(D_j, D_i)}{Cardinality(S)} \tag{5.10}$$

The IDS scores are computed by Eq. 5.2, where D_i is the i^{th} document in a topic. S is the set of all documents other than D_i under the same topic. The IDS score provides the average similarity score for system summaries against the source documents.

• **Redundancy**: It computes the level of information that is being passed from source document to system summaries and is computed using Eq. 5.4,

$$Redundancy(D) = \sum_{\omega_i} S_D(\omega_i).\log(S_D(\omega_i)) \tag{5.11}$$

where S_D represents the probability distribution of a system summary D. ω_i represents the $\tilde{i}th$ semantic unit in the distribution.

Chapter 6

Experiments and Results

6.1 Experiments

6.1.1 Oracle Summaries

Similar to [15], we rank sentences based on their overlap with the reference summary w.r.t. ROUGE-1 and ROUGE-2 scores. We select first k words for the oracle summary, where k is the maximum word count amongst reference summaries of a corpus.

6.1.2 Corpus-related experiments

Position Bias: We use Eq. 5.5 to calculate the relative importance of segments in candidate documents. We use BERT embeddings for vector representations and the cosine similarity metric to compare importance amongst 3 segments of each document: first, middle and last. **Inter Document Similarity (IDS) and Redundancy:** For IDS (Eq. 5.2) and Redundancy (Eq. 5.4), we normalize values to bring them between 0 and 1. We use LDA to obtain topical model distributions and decide on the number of topics based on Coherence score as shown in Equation 6.1.

$$CoherenceScore = \sum_{i<j} \Big(\log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \Big) \tag{6.1}$$

where w_i and w_j are the i^{th} and j^{th} words among the top words. $p(w)$ represents the probability of a word in a random document and $p(w_i, w_j)$ represents the probability of w_i and w_j occuring together in a random document.

Pyramid Score: We use PyrEval [10] to generate the summarization content units (SCUs, Section 5.1) and obtain the pyramid score for each summarization instance. We average this score across the dataset to report results.

6.1.3 System-related experiments

Systems are trained and inferences are generated with the following hyperparameter settings. **Extractive System Settings:** LexRank, TextRank, MMR and ICSISumm summaries are truncated in accordance to corpus reference summary lengths. For MMR, the value of λ is set to 0.5. To generate summaries using ICSISumm, minimum sentence length is set to 10 words with the R_C and R_R parameters set as $True$, which are flags for removing citations and removing redundancy respectively.

Abstractive System Settings: We train the abstractive systems using standard hyperparameters as proposed in the original works. We use models trained on the CNN/Dailymail corpora to evaluate on the DUC, TAC corpus. We use models trained on Multinews to evaluate summaries on the Opinosis and Multinews. Since CQASumm is significantly different from the other corpora, we train, validate and test models on the same corpus using a 80-10-10 split. **ROUGE Scores:** ROUGE-1 (R1), ROUGE-2 (R2) and ROUGE-L (RL) scores are computed between reference summaries and system generated summaries using scripts by [5]. **F1 Score:** We compute the word level F1 Score between a system generated summary and the oracle summary for that summarization instance. **Summary Abstractness:** We determine the extent of newly formed tokens that are non-derivative of source texts by computing higher order n-gram overlap percentage (Section 5.2). **Position Distribution:** Similar to corpus position bias, position distribution for systems is calculated by segmenting the source document into 3 parts and finding the similarity of the generated sentences to different parts of the candidate documents (Section 5.2). **IDS and Redundancy:** Topical model distribution is obtained using LDA,. IDS and redundancy are computed using Eq. 5.2 and Eq. 5.4 respectively; the obtained scores are normalized between [0,1].

6.2 Results

With respect to **Position bias**, news datasets i.e., DUC, TAC and Multinews exhibit 5.4% shift in importance on average in the candidate documents between the most and the least important

segment respectively as shown in Figure 6.1. CQASumm, on the other hand, only shows 1.4% change in importance between the most and least important segment.**A lower shift tend to indicate that while generating the reference summary, all segments have been given similar consideration within any 3-fold segmented article**.

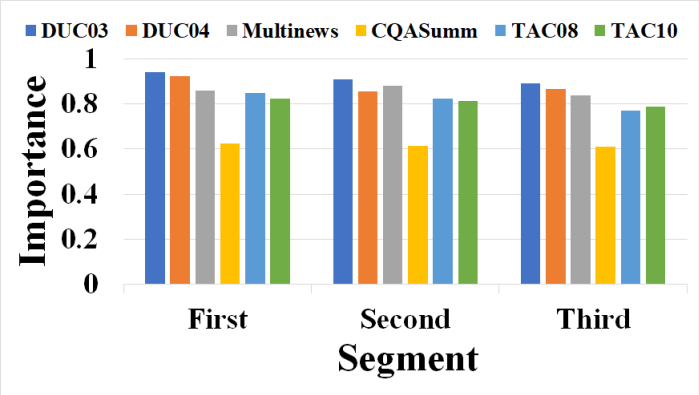
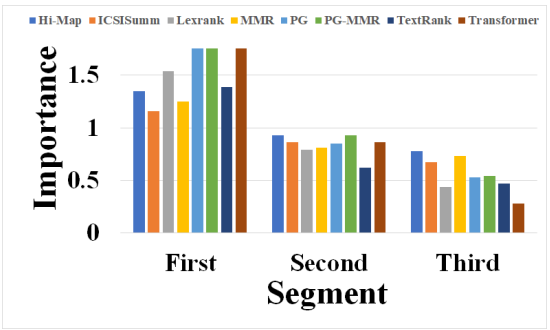
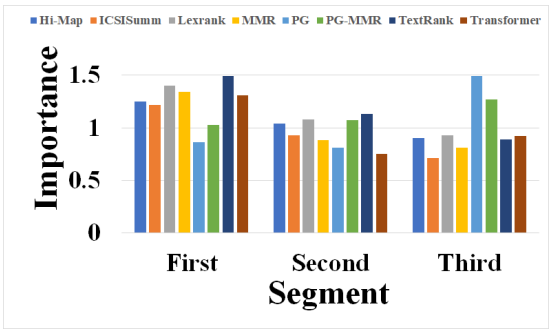


Figure 6.1: Position Bias in Corpus

For **Abstractness**, CQASumm and Multinews respectively have 41.41% and 32.28% novel unigrams in the reference summary. On the contrary, DUC, TAC and Opiniosis reference summaries contain only 11% new unigrams as shown in Table 6.1. **A high Abstractness score highlights the presence of more distinctive phrases in reference summary**. In terms of **Redundancy**, news datasets like Multinews, DUC and TAC exhibit a skewness of -0.40533 in terms of topical content on average. Opinosis and CQASumm show a redundancy value of -0.061 on average. **The farther this score is from 0, the better a document is distributed over its semantic units in the distribution, hence lesser the redundancy**. As evident from Equation 5.4, redundancy is maximized if all semantics units have equal distribution i.e., $P(\omega_i) = P(\omega_j)$. For **Inter Document similarity**, Multinews shows an overlap score of -1.03 which is significantly higher inter document content overlap as compared to DUC, TAC, CQA-



(a) Position Bias in systems



(b) Position Bias in systems with jumbled senetences

Figure 6.2: Position Bias across Systems

Summ and Opinosis whose combined average is -6.433. **The farther this score is from 0, the lesser inter document overlap there is in terms of semantic unit distribution**. As shown in Equation 5.2, the numerator calculates relevance which can be interpreted as the average surprise of observing one distribution while expecting another. This score is small if the distributions are similar i.e., $P_A \approx P_B$ from Equation 5.3.

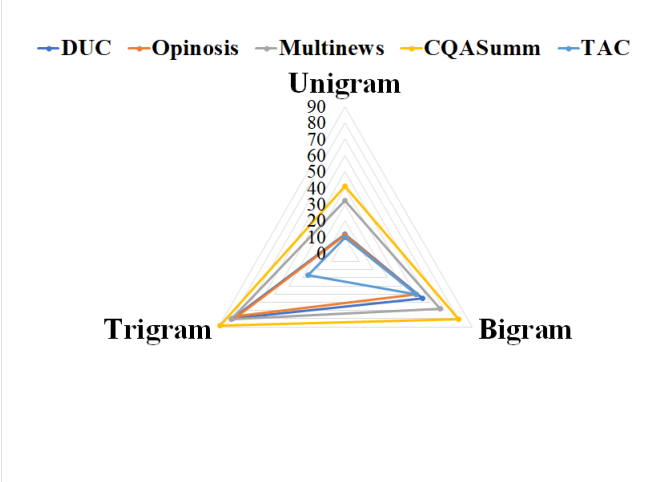
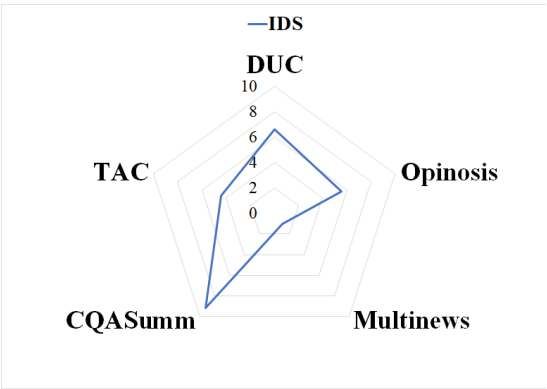
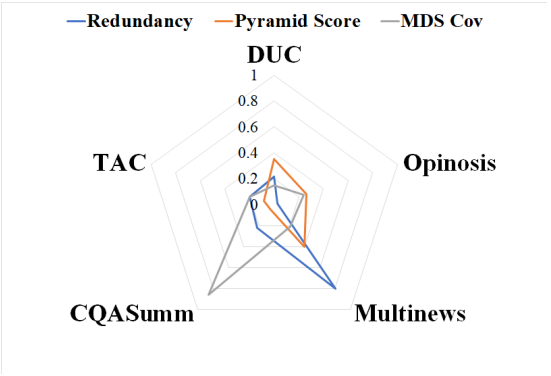


Figure 6.3: Abstractness across varios corpora

For reference quality, in terms of coverage of important aspects of documents **Pyramid Score** shows DUC, TAC, Multinews and Opinosis to be of high quality in Table 6.1 with an average of 0.3325. CQASumm fares last in terms of covering important content across all candidate documents with a Pyramid score of 0.05. **Higher Pyramid score values indicate that reference summary covers the SCUs at the top of the pyramid better**. SCUs present at the top are the ones occurring in most articles and thus can deemed as important. In terms of **Multi-document Summarization Coverage** shows CQASumm to be the dataset with the



(a) IDS across various corpora



(b) Pyramid Score, Redundancy and MDS Coverage across various corpora

Figure 6.4: Corpus Metrics across various corpora

highest propensity of variation in similarity between document and summary at 0.086. Other datasets exhibit a variation of 0.015 on average. **Higher MDS Cov scores suggest the difference in importance given to each document while generating the summary is higher**. As evident from Equation 5.6, Variance across the similarities is high if the similarity scores across the document-summary pairs are highly uneven. For inter-dependency between corpus metrics, our analysis shows Inter Document similarity and Pyramid Score to be strongly positively correlated with a coefficient of 0.8296 across the datasets, while for Redundancy and Inter Document Similarity, the coefficient jumps to the other end with a value of -0.8454 indicating strong negative correlation.

In terms of ROUGE, Extractive summarization systems on an average achieve higher ROUGE score of up to 10% on R1 compared to Abstractive systems as shown in Table 6.2, this tendency is observed in both R1 and R2 scores. **F1 score** follows a similar trend in Table 6.2, Extractive systems score 30% higher than Abstractive systems on an average. With respect to **Abstractness**, Transformer summaries exhibit the highest novel unigram formation of 31.9% on multinews and maintains an average of 28.5% across PG, PG-MMR and Hi-Map systems. DUC, TAC and Opinosis summaries from Pointer Generator models shows an average novel unigram formation of just 1.8%, 2% and 4% respectively, on the other hand transformers and Hi-Map on DUC, TAC and Opinosis achieves an average of 11.5%, 11.8% and 6.45% respectively. In terms of **Inter document similarity**, LexRank summaries for TAC shows higher content overlap by 30% compared to DUC, Opinosis, Multinews and CQASumm. Systems such as MMR, ICSISumm, PG, PG-MMR, Transformer and Hi-Map shows minimal deviation in the IDS scores across corporas. Overall, Transformers and Hi-map exhibits a higher content overlap by 35% compared to ICSISumm. For **Redundancy**, TextRank summaries on TAC attain a score of -1.53, while MMR on TAC have least topic coverage with an average score of -0.068. ICSISumm infers the highest topical coverage by 34% compared to Transformer that infers the least topical coverage among all corpora.

Table 6.1: Corpus metrics showing Abstractness, Redundancy (Red), Inter Document Similarity (IDS), Pyramid Score (Pyr) and MDS Coverage (Cov).

Dataset	**Metric**						
	Abstractness			**Red**	**IDS**	**Pyr**	**Cov**
	1-gram	**2-gram**	**3-gram**				
DUC	11.50	54.66	79.29	-0.21	-6.60	0.35	0.01
Opiniosis	11.5	50.36	76.31	-0.02	-5.53	0.26	0.02
Multinews	32.28	67.53	80.45	-0.80	-1.03	0.4	0.02
CQASumm	41.41	80.72	88.79	-0.22	-9.16	0.05	0.08
TAC	9.91	50.26	26.17	-0.19	-4.43	0.32	0.02

Table 6.2: Metric (Met) showing ROUGE Scores (R1, R2), F1 Score (F1) with oracle summaries, Abstractness (Abs.), Redundancy (Red.) and Inter Document Similarity (IDS) of system summaries for DUC, TAC, Opinosis (Op), Multinews (Multin) and CQASumm (CQAS).

System	Met	Dataset				
		DUC	**TAC**	**Opino**	**Multin**	**CQAS**
LexRank	R1	35.56	33.1	33.41	38.27	32.22
	R2	7.87	7.5	9.61	12.7	5.84
	F1	31.34	31.51	31.05	41.01	49.71
	Red.	-0.136	**-0.104**	-0.278	-0.29	**-0.364**
	IDS	**-3.377**	**-1.87**	-3.526	-2.53	-2.17
TextRank	R1	33.16	44.98	26.97	38.44	28.94
	R2	6.13	9.28	6.99	13.1	5.65
	F1	40.8	29.69	31	38.44	46.3
	Red.	-0.25	-1.553	-0.342	-0.208	-0.247
	IDS	**-0.196**	**-5.97**	-2.745	-1.879	-2.137
MMR	R1	30.14	30.54	30.24	38.77	29.33
	R2	4.55	4.04	7.67	11.98	4.99
	F1	30.57	28.3	31.8	42.07	45.48
	Red.	-0.266	-0.068	-0.255	-0.17	-0.288
	IDS	-2.689	-2.135	**-3.213**	**-1.83**	-2.059
ICSI-Summ	R1	**37.31**	28.09	**27.63**	37.2	28.99
	R2	9.36	3.78	5.32	13.5	4.24
	F1	**24.27**	27.82	29.83	**44.71**	50.98
	Red.	-0.327	-0.283	-0.328	-0.31	-0.269
	IDS	-3.357	-1.903	-3.244	-3.14	-2.466
PG	R1	31.43	31.44	**19.65**	**41.85**	35.09
	R2	6.03	6.4	1.29	12.91	5.52
	F1	23.08	26.32	16.08	43.89	21.85
	Red.	-0.16	-0.254	-0.188	-0.28	-0.12
	IDS	-2.1	**-1.93**	-2.1	-2.103	**-0.5**
	Abs.	1.7%	**1.0**%	4.0%	**28.0**%	6.5%
PG-MMR	R1	36.42	40.44	19.8	40.55	36.54
	R2	9.36	14.93	1.34	12.36	6.67
	F1	24.3	26.9	**16.39**	**43.93**	21.72
	Red.	-0.17	-0.26	-0.172	-0.29	-0.142
	IDS	-2.4	-1.87	-1.9	-1.98	-0.72
	Abs.	1.9%	2.0%	4.0%	27.5%	6.9%
CopyTransformer	R1	28.54	31.54	**20.46**	**43.57**	30.12
	R2	6.38	5.9	1.41	14.03	4.36
	F1	**15.72**	17.82	16.38	**44.54**	21.35
	Red.	**-0.177**	-0.17	-0.189	-0.18	**-0.273**
	IDS	-1.914	-1.867	-1.589	-1.89	-2.239
	Abs.	9.0%	9.0%	4.9%	31.9%	9.2%
Hi-Map	R1	35.78	29.31	18.02	43.47	31.41
	R2	8.9	4.61	1.46	14.89	4.69
	F1	25.89	24.3	20.36	42.55	19.84
	Red.	-0.172	-0.2	-0.16	-0.23	-0.26
	IDS	-1.62	-1.652	-1.8	-1.788	-2.223
	Abs.	14.0%	14.7%	**8.0**%	**26.7**%	7.0%

Chapter 7

Inferences and Conclusion

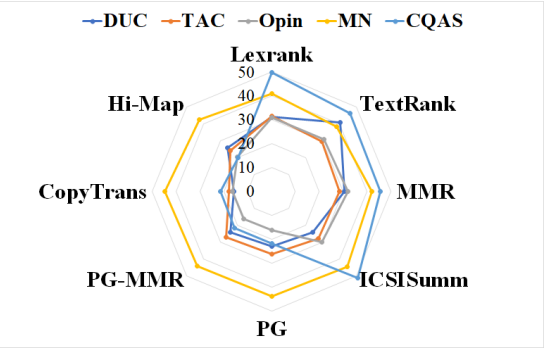
7.1 Inferences

7.1.1 Corpus Metric Inferences

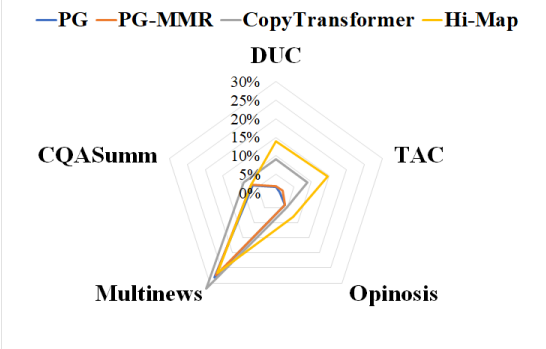
We base our inferences on the experiments performed by us, described in Chapter 6.

Dataset Position Bias: News articles tend to bear important summary information in the opening lines. This is corroborated with high Lead-3 scores in most news corpora. This behaviour is not prevalent in CQA or opinion corpus data where the important phrases are spatially spread out across text. These position bias have a crucial role in determining the performance of summarization systems on a corpora.

Reference Abstractness: Previous MDS corpora like DUC, TAC and Opinosis have low



(a) F1 Score across various systems



(b) Abstractness across various systems

Figure 7.1: Abstractness and F1 across various systems

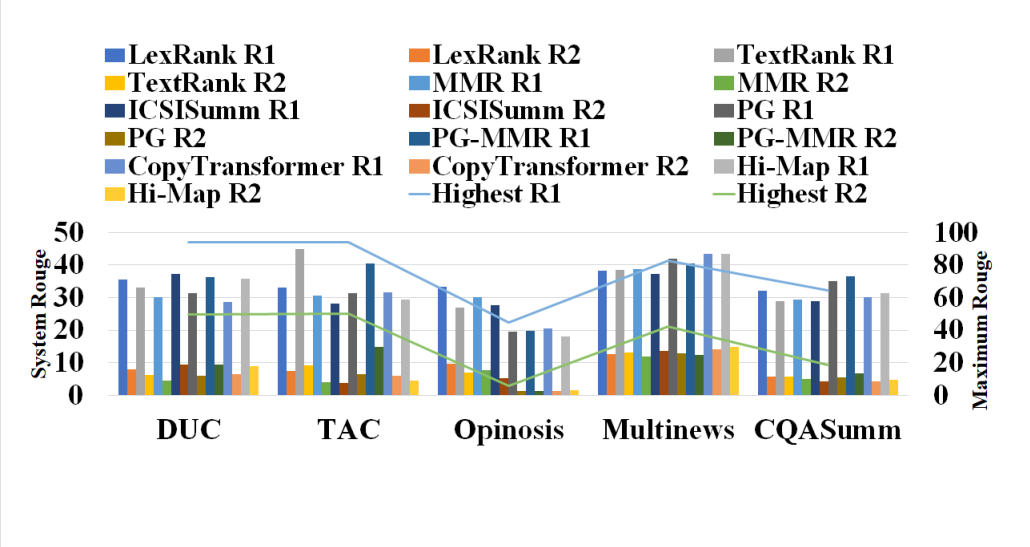


Figure 7.2: Maximum Rouge across various systems

corpus size as well as abstractness. However news corpora like Multinews are significantly larger in size and more abstract. CQASumm is highly abstract because the views chosen to form the best answer are an individual's point of view and are not likely to significantly match with another user's views.

Dataset Redundancy: Redundancy is a measure of importance distribution between different topics discussed in a summarization instance. As shown in Table 6.1, candidate documents in Multinews, DUC and TAC revolve around a single key entity of that topic like a person, a place, an event etc. related to the topic. Crowdsourced opinion based datasets like CQASumm and Opinosis have more breadth and show versatility in terms of topical content.

Inter Document Similarity (IDS): Table 6.1 implies that corpora with IDS values closer to 0 have higher content overlap between candidates documents as compared to the ones farther from 0. Multinews has the highest Inter Document Similarity amongst the studied datasets, while CQASumm has the lowest value. Documents in rest of the datasets show moderate overlap. It is important to understand how the overlap within documents leads to performance of summarization systems.

Corpus Pyramid Score: We use pyramid score to determine if the important and redundant topics in candidate documents have been elected to be part of the reference summary. We believe this is an important metric to evaluate the quality of a corpora or compare the work of different annotators within a corpora. While DUC, TAC, Multinews and Opinosis have significantly high Pyramid scores, we find it considerably low in the web generated corpora such as CQASumm. This raises questions on the quality of large crowdsourced corpora and their applications as

training data for smaller corpora. We infer that the datasets with higher degree of document overlap tend to have references which cover essential content better, while the redundancy trend shows that with higher overlap documents tend to be less redundant and more versatile.

Multi-document Summarization Coverage (MDS Cov) can also be used to measure the quality of dataset reference or to compare between annotators of a summarization corpora. We use MDS Cov to indicate the extent to which the set of candidate documents under a topic is covered to highlight the very definition of multi-document summarization. Corpora curated with human annotator based references fare better compared to crowdsourced ones. CQASumm having significantly higher MDS Cov indicates that most of the time *best answers* in CQA threads are not a proper reflection of the rest of the answers and are unsuitable as a reference in majority of the dataset.

7.1.2 System Metric Inferences

Inferences learnt from system summaries reflect how summarization systems are influenced by the metrics of the corpora on which they are being trained/tested.

• MDS systems under consideration are ranked differently in terms of ROUGE on different corpora; leading to a dilemma whether to declare a system superior to others without testing on all types of datasets (Fig. 7.2 and Table 6.2).

• Extractive MDS systems under consideration outperform abstractive summarization systems by up to 10% on ROUGE-1 and up to 30% on F1 Scores, showing contradictory behavior in comparison to single-document summarization systems where state-of-the-art abstractive systems are known to outperform the former (Fig. 7.2 & Fig. 7.1a).

• The best summarization system on each corpus obtains a score $39.6\%, 47.8\%, 75.02\%, 54.5\%, 49.9\%$ of the oracle upper bound on DUC, TAC, Opinosis, Multinews and CQASumm respectively, indicating that summarization on Opinosis and Multinews is partially solved problem; while DUC, TAC and CQASumm exhibit considerable scope for improvement (Fig. 7.2).

• Hi-Map and CopyTransformer generate more abstract summaries (17.5% and 16% novel unigrams respectively) in comparison to PG and PG-MMR (Fig. 7.1b).

• Averaging over systems and comparing corpora, we notice that Multinews and CQASumm achieve the highest abstractness (27% and 7% respectively), which might be a result of these two corpora having the most abstract reference summaries.

• Abstractive systems exhibit a 55% shift in importance between the first and the second seg-

ments of generated summaries, whereas extractive systems show an average shift of only 40%, implying that abstractive systems have a stronger tendency to display layout bias (Fig. 6.2a and Fig. 6.2b).

• While DUC, TAC and Opinosis summaries generated from PG trained models exhibit lower novel unigrams formation, the same for CopyTransformer and Hi-Map on DUC, TAC and Opinosis shows a higher unigram formation on average (Fig. 7.1b).

• In terms of Inter Document Distribution, LexRank summary for TAC and CQASumm shows more variance across documents compared to DUC, Opinosis and Multinews. TextRank summary on DUC, TAC and CQASumm, MMR summary on DUC, and Hi-Map summary on CQA-Summ show higher variances as well. Systems such as PG, PG-MMR and CopyTransformer show minimal deviation in the document participation across corpora.

• In terms of Topic Coverage, extractive systems show better coverage than abstractive systems (Table 6.2), which might be a result of extractive systems being based on sentence similarity algorithms which find important sentences, reduce redundancy and increase the spread of information from different segments of the candidate document.

Table 7.1: Pearson correlation between corpus and system with column 4 (**First**) between Abstractness of corpora and system, column 5 (**Second**) between Abstractness of corpora and ROUGE-1 score of systems across datasets and column 6 (**Third**) showing Position Bias correlation between system and corpora.

System	**Metric**				
			Position correlation		
	Abs. corr	**R-1 corr**	**First**	**Second**	**Third**
LexRank	-	0.08	0.88	0.06	0.96
TextRank	-	-0.24	0.91	0.76	0.97
MMR	-	0.32	0.86	0.09	0.97
ICSISumm	-	0.11	0.39	0.53	0.72
PG	0.57	0.65	0.80	-0.80	-0.98
PG-MMR	0.57	0.33	0.84	-0.69	-0.91
CopyTrans	0.47	0.50	0.84	-0.31	-0.79
Hi - Map	0.11	0.45	0.74	-0.11	-0.46

7.1.3 Discussion on Research Question

Q1. How should one model the quality of an MDS corpus as a function of its intrinsic metrics? What guidelines should be followed to propose MDS corpora for enabling a fair comparison with existing datasets? The quality of an MDS corpus is a function of two independent variables: the *quality of the candidate documents* and the *quality*

of the reference summary. Our findings suggest that a framework for future MDS datasets should provide scores measuring their standing w.r.t. both the above factors. The former is usually crowd-source dependent, while the latter is usually annotator dependent. While Inter Document Similarity, Redundancy, Layout Bias and Inverse-Pyramid Score are indicators of the properties of the candidate document, metrics such as Abstractness of the reference summary and Pyramid Score are ground-truth properties. While all these metrics should be reported by imminent corpora proposers to enable comparisons with existing corpora and systems, we feel that the average *Pyramid Score* and *Inverse-Pyramid Score* must be reported as they are strong indicators of generic corpus quality. Other metrics such as IDS, Redundancy, Abstractness etc. can be modeled according to task-based requirements. These classifications can find applications in a range of NLP tasks, e.g., in choosing a training dataset that is closest to a small test dataset.

Q2. Why do the ROUGE-based ranks of different MDS systems differ across corpora? How should an MDS system which is to achieve reasonably good ROUGE score on* all *corpora look like? From Table 6.2 within studied systems, in terms of ROUGE-1, ICSISumm achieves the best score on DUC, TextRank on TAC, LexRank on Opinosis, CopyTransformer on Multinews and LexRank on CQASumm. Hence as of today, no summarization system strictly outperforms others on every corpus. We also see that CopyTransformer which achieves state-of-the-art performance on Multinews, achieves 10 points less than the best system on DUC. Similarly, LexRank, the state-of-the-art performer on CQASumm, achieves almost 12 points less than the best system on TAC. Therefore, a system that performs reasonably well across all corpora, is also missing. This is because *different corpora are high on various bias metrics, and summarization systems designed for a particular corpus take advantage and even aggravate these bias.* For example, summarization systems proposed on news based corpora are known to feed only the first few hundred tokens to neural models, thus taking advantage of the layout bias. Feeding entire documents to these networks has shown to relatively lower performance. Systems such as LexRank are known to perform well on candidate documents with high inter-document similarity (e.g., Opinosis). Solving summarization problem for an unbiased corpus is a harder problem, and for a system to be able to perform reasonably well on any test set, it should be optimized to work on such corpora.

Q3. Why do systems show bias on different metrics, and which other system and corpus attributes are the reason behind it? We begin by studying how *abstractness of generated summaries is related to the abstractness of corpora the system is trained on.* For this, we calculate the Pearson correlation coefficient between the abstractness of generated sum-

maries and references across different datasets. From Table 7.1, we infer that PG, PG-MMR and CopyTransformer show a positive correlation which implies that they are *likely to generate more abstract summaries if the datasets on which they are trained have more abstract references.* Lastly, we infer *how Layout Bias in system-generated summaries is dependent on the layout bias of reference summaries.* The last three highlighted columns of Table 7.1 infer that the abstractive systems such as PG, PG-MMR, Hi-Map and CopyTransformer show a high negative correlation for the end segments while maintaining a strongly positive one with the starting segment. On the other hand, extractive systems such as LexRank, TextRank, MMR and ICSISumm maintain a strongly positive correlation throughout the segments. On shuffling the source segments internally, we observe that extractive systems tend to retain their correlation with corpora while abstractive systems show no correlation at all (Fig. 6.2), proving that in supervised systems, *the layout bias in system summaries propagates from the layout bias present in corpora.*

Q4. Is the task of MDS almost solved, or there is still plenty of scope remaining for improvement? In the previous sections, we computed the oracle extractive upper bound summary using greedy approaches to find the summary that obtains the highest ROUGE score given the candidate documents and references. We observe that the best summarization system on each corpus today obtains a score which is 39.6% of the extractive oracle upper bound on DUC, 47.8% on TAC, 75.02% on Opinosis, 54.5% on Multinews and 49.9% on CQASumm. This shows that there is a enough scope for MDS systems to achieve double the ROUGE scores obtained by the best system till date on each corpus except Opinosis. Therefore, we believe that the task of MDS is only partially solved and considerable efforts need to be devoted to improve the systems.

7.2 Conclusion

In this paper, we aimed to study the heterogeneous task of multi-document summarization. We analyzed interactions between widely used corpora and several state-of-the-art systems to arrive at a line of conclusions. We defined MDS as a mapping from a set of non-independent candidate documents to a synopsis that covers *important* and *redundant* content present in the source. We proposed intrinsic metrics to model the quality of an MDS corpus and introduced a framework for future researches to consider while proposing a new corpus. We analyzed how ROUGE ranks of different systems vary differently on different corpora and described what a system that achieves reasonable performance on all corpora would look like. We evaluated how different systems exhibit bias and how their behavior is influenced by corpus properties.

Chapter 8

Future Work

Apart from an attempt to answer, some of the most profound correlations in Multi-Document Summarization, this work raises some questions which might be the basis for further inspection in the field. It is important to understand how the overlap within documents leads to performance of summarization systems. It would be further interesting to model instances of conflict within candidate documents and observe how different summarization systems deal with it. A more detailed outline of corpus properties which might be the core reason for performance across various summarization systems need to drawn.

Bibliography

[1] Benikova, D., Mieskes, M., Meyer, C. M., and Gurevych, I. Bridging the gap between extractive and abstractive summaries: Creation and evaluation of coherent extracts from heterogeneous sources. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (2016), pp. 1039–1050.

[2] Brin, S., and Page, L. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International Conference on World Wide Web 7* (1998), pp. 107–117.

[3] Carbinell, J., and Goldstein, J. The use of mmr, diversity-based reranking for reordering documents and producing summaries. *SIGIR Forum 51*, 2 (Aug. 2017), 209–210.

[4] Chowdhury, T., and Chakraborty, T. Cqasumm: Building references for community question answering summarization corpora. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data* (2019), ACM, pp. 18–26.

[5] Diego999. Diego999/py-rouge, Jul 2019.

[6] DUC. Document Understanding Conferences. [online] Available at: https://duc.nist.gov/.

[7] Erkan, G., and Radev, D. R. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research 22* (Dec 2004), 457–479.

[8] Fabbri, A. R., Li, I., She, T., Li, S., and Radev, D. R. Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749* (2019).

[9] Ganesan, K., Zhai, C., and Han, J. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Con-*

ference on Computational Linguistics (2010), Association for Computational Linguistics, pp. 340–348.

[10] Gao, Y., Warner, A., and Passonneau, R. Pyreval: An automated method for summary content analysis. In *LREC 2018 - 11th International Conference on Language Resources and Evaluation* (1 2019), H. Isahara, B. Maegaard, S. Piperidis, C. Cieri, T. Declerck, K. Hasida, H. Mazo, K. Choukri, S. Goggi, J. Mariani, A. Moreno, N. Calzolari, J. Odijk, and T. Tokunaga, Eds., LREC 2018 - 11th International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA), pp. 3234–3239.

[11] Gehrmann, S., Deng, Y., and Rush, A. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels, Belgium, Oct.-Nov. 2018), Association for Computational Linguistics, pp. 4098–4109.

[12] Gillick, D., Favre, B., and Hakkani-Tür, D. The icsi summarization system at tac 2008. In *Tac* (2008).

[13] Grusky, M., Naaman, M., and Artzi, Y. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (New Orleans, Louisiana, June 2018), Association for Computational Linguistics, pp. 708–719.

[14] Hirao, T., Fukusima, T., Okumura, M., Nobata, C., and Nanba, H. Corpus and evaluation measures for multiple document summarization with multiple sources. In *Proceedings of the 20th international conference on Computational Linguistics* (2004), Association for Computational Linguistics, p. 535.

[15] Hirao, T., Nishino, M., Suzuki, J., and Nagata, M. Enumeration of extractive oracle summaries. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (Valencia, Spain, Apr. 2017), Association for Computational Linguistics, pp. 386–396.

[16] Jung, T., Kang, D., Mentch, L., and Hovy, E. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. In *Proceedings of the 2019*

Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (Hong Kong, China, Nov. 2019), Association for Computational Linguistics, pp. 3324–3335.

[17] Kryscinski, W., Keskar, N. S., McCann, B., Xiong, C., and Socher, R. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (Hong Kong, China, Nov. 2019), Association for Computational Linguistics, pp. 540–551.

[18] Lebanoff, L., Song, K., and Liu, F. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels, Belgium, Oct.-Nov. 2018), Association for Computational Linguistics, pp. 4131–4141.

[19] Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* (Barcelona, Spain, July 2004), Association for Computational Linguistics, pp. 74–81.

[20] Mihalcea, R., and Tarau, P. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (Barcelona, Spain, July 2004), Association for Computational Linguistics, pp. 404–411.

[21] Nenkova, A., and Passonneau, R. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004* (Boston, Massachusetts, USA, May 2 - May 7 2004), Association for Computational Linguistics, pp. 145–152.

[22] Peyrard, M. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019), pp. 1059–1073.

[23] Peyrard, M. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019), pp. 5093–5100.

[24] See, A., Liu, P. J., and Manning, C. D. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vancouver, Canada, July 2017), Association for Computational Linguistics, pp. 1073–1083.

[25] TAC. Text Analysis Conferences. [online] Available at: https://tac.nist.gov/.

[26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *CoRR abs/1706.03762* (2017).

[27] Zhu, C., Yang, Z., Gmyr, R., Zeng, M., and Huang, X. Make lead bias in your favor: A simple and effective method for news summarization, 2019.