Query-Focused Extractive Summarization for Sentiment Explanation

by

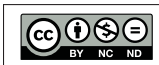
Ahmed MOUBTAHIJ

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT OF A MASTER'S DEGREE
WITH THESIS
M.A.Sc.

MONTREAL, APRIL 26, 2023

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

Ahmed Moubtahij, 2023



This Creative Commons license allows readers to download this work and share it with others as long as the author is credited. The content of this work cannot be modified in any way or used commercially.

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mrs. Sylvie Ratté, Thesis supervisor
Department of Software and Information Technology Engineering, École de Technologie Supérieure

Mr. Yazid Attabi, Thesis Co-Supervisor
Croesus

M. Tony Wong, Chair, Board of Examiners
Departement of Systems Engineering, École de Technologie Supérieure

M. Luc Duong, Member of the Jury
Department of Software and Information Technology Engineering, École de Technologie Supérieure

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON APRIL 18TH, 2023

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

This work was supported by the Mitacs organization (project reference numbers: IT32729; FR94073; QC-ISDE), Croesus and the Arbour Foundation. I thank Sylvie Ratté, my thesis supervisor, for her continued support and trust in my abilities. I'm also grateful to my supervisors at Croesus for providing the necessary hardware and software infrastructure, and to Stéphane Gazaille for his mentoring in the initial stages of the project.

Génération de résumés extractifs orientés-requête pour l'explication de sentiments

Ahmed MOUBTAHIJ

RÉSUMÉ

L'analyse constructive des sentiments exprimés dans des commentaires de clients requiert typiquement un résumé des causes principales de leur ressenti, et ce, à partir d'une quantité importante de documents textuels. Afin d'automatiser de tels efforts, nous nous appuyons sur la tâche de génération de résumés orientés-requête. Les modèles de génération existant sont souvent entravés par la dissonance linguistique entre la requête et les documents sources. Dans ce mémoire, nous proposons et concrétisons un cadre multi-biais pour parer au problème de dissonance à un niveau générique et indépendant du domaine, puis nous formulons des approches spécialisées pour l'explication du sentiment interrogé, à savoir, des stratégies de biais et d'augmentation de requêtes. Nous obtenons des résultats expérimentaux qui surpassent les modèles de base sur un ensemble-propriétaire de données du monde réel.

Mots-clés: génération de résumés orientés-requête, analyse de sentiment, explication de sentiment, augmentation de requête

Query-Focused Extractive Summarization for Sentiment Explanation

Ahmed MOUBTAHIJ

ABSTRACT

Constructive analysis of clients' feedback often requires determining the cause of their sentiment from a substantial amount of text documents. In order to assist and improve the productivity of such endeavors, we leverage the task of Query-Focused Summarization (QFS). Models of this task are often impeded by the linguistic dissonance between the query and the source document(s). We propose and substantiate a multi-bias framework to help bridge this gap at a domain-agnostic, generic level, then we formulate specialized approaches for the problem of sentiment explanation through sentiment-based biases and query expansion. We achieve experimental results outperforming baseline models on a real-world proprietary sentiment-aware QFS dataset.

Keywords: Query-Focused Summarization, Sentiment Analysis, Sentiment Explanation, Query Expansion

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ACRONYMS

ÉTS	École de Technologie Supérieure
ML	Machine Learning
NN	Neural Network
RNN	Recurrent Neural Network
NLP	Natural Language Processing
QFS	Query-Focused Summarization
QFES	Query-Focused Extractive Summarization
QFES	Query-Focused Abstractive Summarization
ESS	Explicative Sentiment SUmmarization
BERT	Bidirectional Encoders Representations from Transformers
LM	Language Model
MLM	Masked Language Model
MPB	MLM-Pattern Based
MMR	Maximal Marginal Relevance
BQ	Baseline Query
CBFS	Compound Bias-Focused Summarization
ERT	Expanded Reference Terms
FRW	Frequent Reference-Words
FRP	Frequent Reference-Phrases

BTR	Biased TextRank
MBTR	MultiBias TextRank
LR	Language Register
IC	Information Content
DUC	Document Understanding Conference
ABSA	Aspect Based Sentiment Analysis

INTRODUCTION

Sentiment analysis is the Natural Language Processing (NLP) task of predicting the affective state of a text passage. It is generally useful for applications concerned with feedback analysis of experiences (e.g., products, events or services). However, simply being aware of the sentiment does not improve the experience; this purpose requires knowledge of the specific causes and features related to the sentiment.

Given a multitude of documents, a sentiment of interest (e.g. negative or positive), and a query regarding the targeted entities (e.g., a specific product, date, or location), our main objective is to provide an informative summary of the input documents that justifies the queried sentiment. This requirement describes a constrained *Query-Focused Summarization (QFS)* task, which we term as *Explicative Sentiment Summarization (ESS)*. Compared to the Question Answering task's factoid outputs, the QFS task is motivated by more complex and contextually rich responses. Thus, it is a more appropriate parent task for ESS, which consists of elaborating on the cause(s) of the queried sentiment.

A common shortcoming of the QFS task and its proposed models is the putative gap between the source text and the input query in terms of *Language Register* (LR, formality level) and *Information Content* (IC, from Shannon's Information Theory). An LR gap occurs when, e.g., a colloquial query formulation addresses source text written in formal style or in domain-specific terminology. An IC gap is typically incurred by the generic semantic coverage of short queries in relation to the specific semantics in detailed source text passages.

Our aforementionned main objective of explicative sentiment summarization factors in the issues of IC and LR dissonance between the query and the source text.

Our following contributions first address the linguistic dissonance issue in QFS at a generic level, then at a specialized level for our purpose of sentiment explanation:

1. We introduce the *Compound Bias-Focused Summarization (CBFS)* (6.4.1) framework to improve the chances of aligning the user's intent with arbitrary and possibly heterogeneous language registers in source documents by supporting multiple query formulations.
2. We concretize the CBFS framework with our *Multi-Bias TextRank* (6.4.2) model and its *Information Content Regularization* (6.4.3) which guides the QFS process towards the desired level of specificity.
3. We introduce the *Explicative Sentiment Summarization (ESS)* task, (6.4.4) which specializes the QFS task by leveraging prior knowledge in a sentiment explanation setting.
4. We substantiate the ESS task with sentiment-based bias computation (6.4.4.2) and query expansion (6.4.4.3).

CHAPTER 1

LITERATURE REVIEW

An overview of the relevant ML architectures heads the hierarchical structure of this chapter. Once these architectures are established as the NNs underlying the NLP tasks of interest, we address the literature of said tasks, i.e., Automatic Summarization, Query-Focused Summarization, and Query Expansion.

Finally, we explore ways of quantitatively and qualitatively analyzing the presented models' outputs. Such is achieved with a survey of automatic summarization evaluation metrics, as well as with a taxonomy of errors in automatic summarization.

Note that subsections are chronologically ordered in terms of the papers' release years.

1.1 Towards modern summarization models

The applications of the neural NLP field (e.g. translation, auto-completion, **summarization**. . . etc.) rely heavily on the ability to **encode words**. As such, the evolution of neural NLP went hand in hand with the evolution of encoders.

The role of encoders in NLP is to encapsulate the semantics of words within a corresponding numerical representation to enable the mathematical transformations used by a Language Model.

A Language Model's task is predicting the next word or sequence of words, given the previous word or sequence of words. The prediction is typically based on pre-learned occurrence probabilities of neighboring words. It may additionally be conditioned on priors such as semantic salience w.r.t the text and relevance w.r.t to a query. The latter priors constitute the basis of the **query-focused summarization** task introduced in 1.3.

1.1.1 Recurrent Neural Networks (RNNs)

The sequential nature of RNNs (Figure 1.1) lends itself well to predicting the next words based on the previous words, that is, the task of a Language Model (LM).

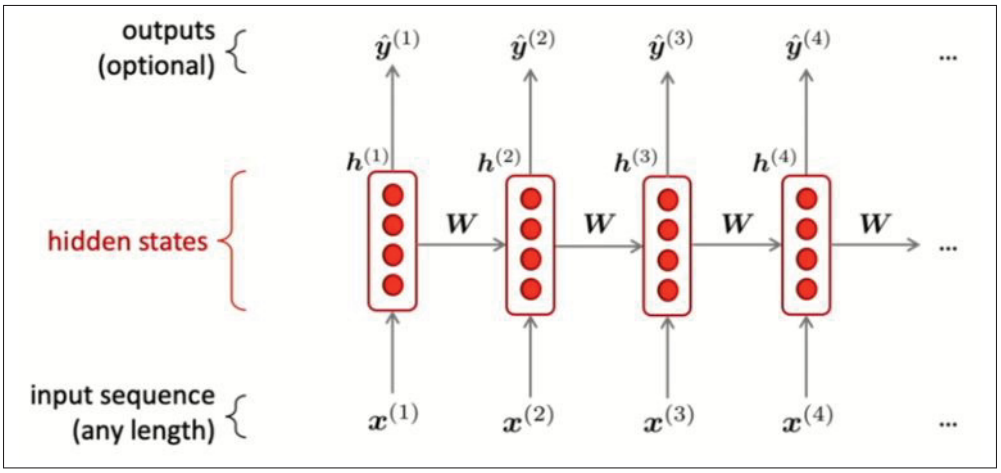


Figure 1.1 RNN structure
Taken from https://www.youtube.com/watch?v=iWea12EAu6U

Hidden states $\boldsymbol{h}^{(t)}$ are processed at each input $\boldsymbol{x}^{(t)}$ where t is the time step. That is, the hidden state mutates over time. The *same* weights matrix $\boldsymbol{W}$ is shared across time steps, which ensures a conjoint training that respects the commonalities between inputs (e.g. words from an observed context) across time steps. $\hat{\boldsymbol{y}}^{(t)}$ are intermediate outputs (required for computing the loss at step t).

1.1.1.1 RNN computation

In Figure 1.2, the input embedding $\boldsymbol{e}^{(t)}$ is expressed as follows:

$$\boldsymbol{e}^{(t)} = \boldsymbol{E}\boldsymbol{x}^{(t)} \tag{1.1}$$

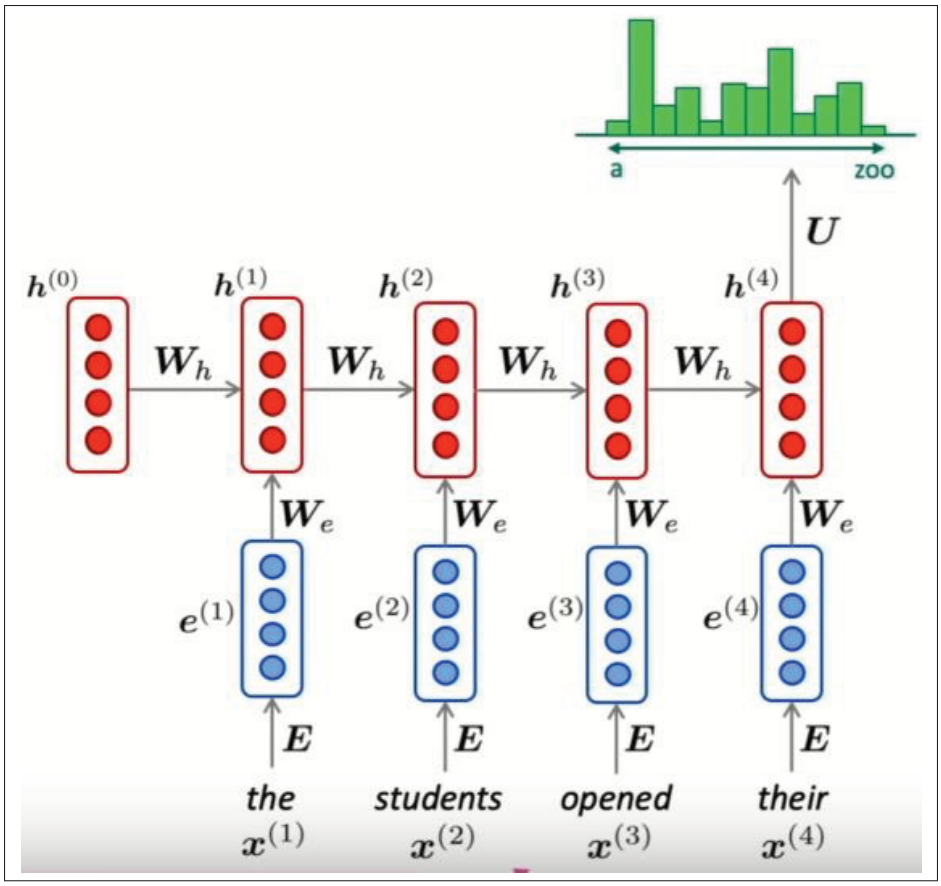


Figure 1.2 RNN Computation
Taken from https://www.youtube.com/watch?v=iWea12EAu6U

where $\boldsymbol{x}^{(t)}$, the one-hot encoded vector per word, is multiplied with the embedding matrix $\boldsymbol{E}$ to produce the input embeddings $\boldsymbol{e}^{(t)}$. The one-hot encoding acts as a $\boldsymbol{E}$ column selector. $\boldsymbol{E}$ can be learned from scratch, or pre-trained and/or fine-tuned.

$$\boldsymbol{h}^{(t)} = \sigma(\boldsymbol{W}_h \boldsymbol{h}^{(t-1)} + \boldsymbol{W}_e \boldsymbol{e}^{(t)} + \boldsymbol{b}_1) \tag{1.2}$$

The hidden state $\boldsymbol{h}^{(t)}$ (Equation 1.2) depends on the previous hidden state $\boldsymbol{h}^{(t-1)}$ and the current input $\boldsymbol{e}^{(t)}$. Both parameters are linearly transformed, respectively with the $\boldsymbol{W}_h$ and $\boldsymbol{W}_e$ weights, and biased with $\boldsymbol{b}_1$ to enable learning. The expression is wrapped in σ, a non-linear function, for more adaptive fitting.

$$\hat{\boldsymbol{y}}^{(t)} = softmax(\boldsymbol{U}\boldsymbol{h}^{(t)} + \boldsymbol{b}_2) \in \mathbb{R}^{|V|} \tag{1.3}$$

The output $\hat{\boldsymbol{y}}^{(t)}$ (Equation 1.3) computes the probability distribution of $\boldsymbol{h}^{(t)}$ over the vocabulary set, V, in $\mathbb{R}^{|V|}$. $\boldsymbol{h}^{(t)}$ is linearly transformed with the $\boldsymbol{U}$ weights and the $\boldsymbol{b}_2$ bias to enable learning on the computed probability.

1.1.1.2 Advantages

- **Independent of input length**: the same weights matrix, $\boldsymbol{W}$, is shared across all time steps, thus, its size does not depend on the sequence length but rather on the embedding dimensionality, which is fixed and known in advance;

  Prior to RNNs, LMs relied on fixed windows when considering the previous words for next word prediction. In fixed-window LMs, the size of $\boldsymbol{W}$ is dependent on the sequence length; each input has its own weights column in $\boldsymbol{W}$, thus, $\boldsymbol{W}$'s upper bound size can only be guessed and not adaptive to an arbitrary sequence length;
- **Better context capture**: in theory, the prediction is informed by deeper past steps. Indeed, given the improved input length capacity, long-term dependencies have a better chance of being considered for next word prediction;
- **Symmetric inputs processing**: the *same* weights matrix $\boldsymbol{W}$ is shared across time steps, which ensures a conjoint training that respects the commonalities between inputs (e.g., words from an observed context) across time steps.

1.1.1.3 Disadvantages

- **Latency in recurrent computation**: computing step n requires first computing step $n-1$. In other words, parallelization is not an option;
- **Long-term memory loss**: in practice, information from many steps back can be lost. This is known as the vanishing gradient problem.

1.1.2 Long Short Term Memory (LSTM)

Hochreiter & Schmidhuber (1997) introduced the LSTM network (Figure 1.3) as a specialized RNN capable of capturing long-term dependencies. Thus, the process of encoding a text passage produces more informed outputs.

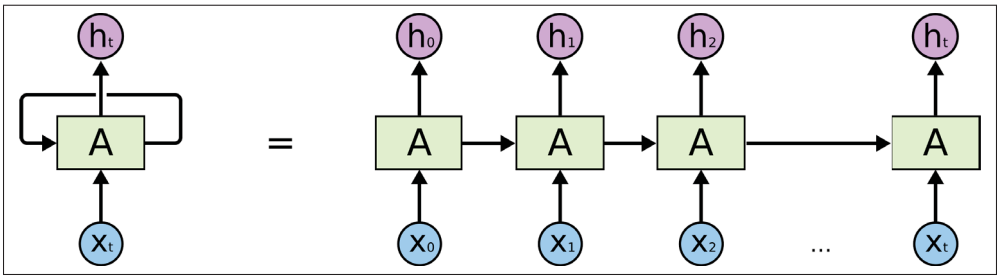


Figure 1.3 Understanding LSTM Networks
Taken from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

LSTM cells (Figure 1.4) incorporate gating logics that filter out irrelevant information and carry over the relevant parts, thus addressing the vanishing gradient problem in basic RNNs.

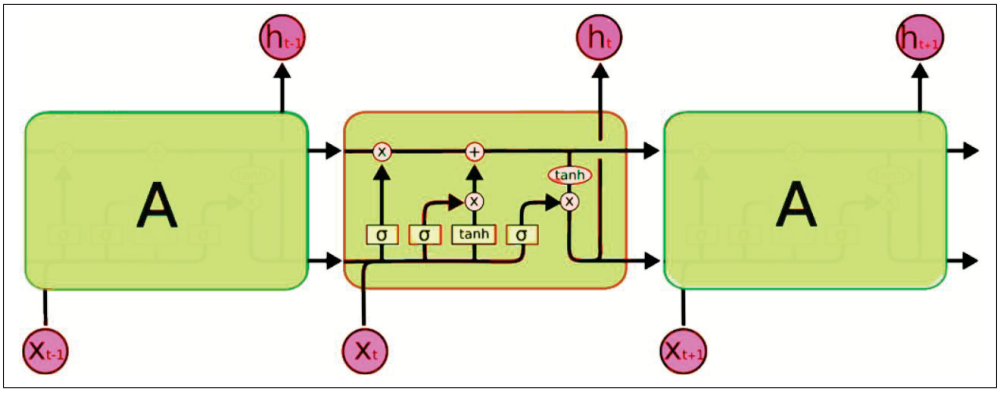


Figure 1.4 Introduction to LSTM
Taken from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

A variant of LSTM, the bidirectional LSTM (Schuster & Paliwal, 1997), performs an additional pass over the processed sequence. It has often been favored in NLP applications for its ability to incorporate information from both past and future dependencies, allowing for a semantic encoding stemming from a richer context.

1.1.3 The Transformer architecture

Bahdanau, Cho & Bengio (2014) introduce the attention mechanism (Figure 1.5), which assigns importance scores to tokens relative to other tokens in a sequence. This approach empowers token representation with contextual awareness. The input tokens are weighted by their relevance to a downstream task (e.g., machine translation, or language modeling), thus improving its performance. Following this work, Vaswani *et al.* (2017) use the attention mechanism in an encoder-decoder-based neural network, i.e., the Transformer deep learning model.

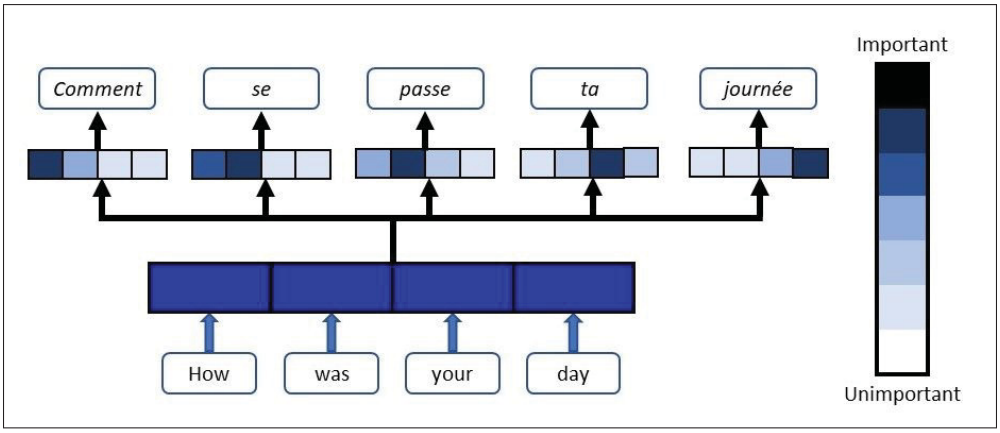


Figure 1.5 Attention mechanism
Taken from https://blog.floydhub.com/attention-mechanism/

Contrary to the recurrent nature of LSTM networks, Transformers are non-sequential. This property enables parallel computation and consequently a substantial reduction in training time. It also allows the attention mechanism in Transformers to process tokens simultaneously, thus eliminating the long-term dependency performance issues found in LSTM networks. Indeed, as Vaswani *et al.* (2017) claim, the attention mechanism allows modeling dependencies regardless of their distance within the input or output sequences.

The original Transformer architecture (Figure 1.6) consists of a stack of encoders connected to one of decoders. Encoders prepare a representation of the input embeddings for the decoders, while the latter produce the desired output. The following describes the main components and operations found in the Transformer architecture:

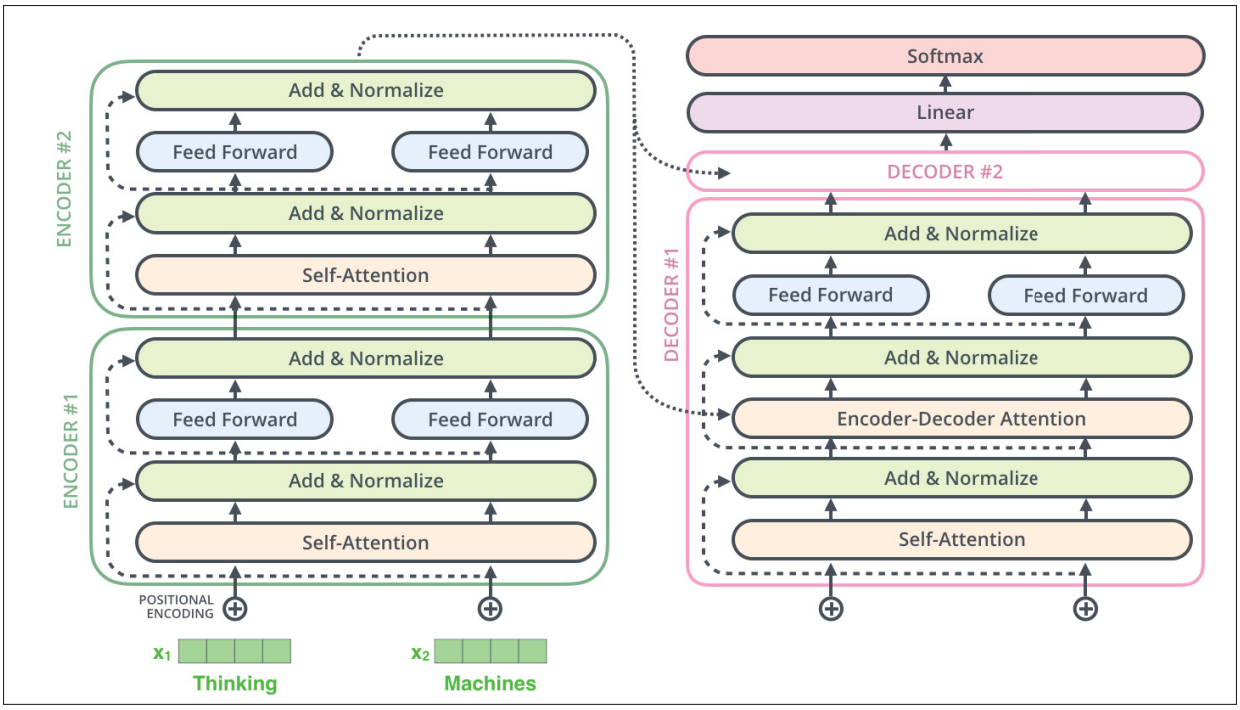


Figure 1.6 The illustrated Transformer
Taken from https://jalammar.github.io/illustrated-transformer/

- **Positional encoding**: language is a syntactic construct, however, the parallel computation of the attention mechanism in the Transformer model renders it permutation-invariant. To fix this discrepancy, the input embeddings are first positionally encoded to preserve order information. This is achieved by summing the embeddings with learned vectors whose role is meaningful distantiation;
- **Add & Normalize layers**: summing the input and its transformation is reminiscent of the residual connections in *ResNet* (He, Zhang, Ren & Sun, 2016) as it similarly solves the vanishing gradient problem. Normalization is technically motivated by the improvement of training time;
- **Self-attention**: the process of weighting the relative importance between token embeddings. These -attention- weights are then used in the weighted sum of the attended token's embedding and its neighbors'. The result is fed to the **Feed-Forward Neural Network (FFNN)** layer to produce the new representation of the attended token;

- **Encoder-decoder attention**: enables an appropriate focus partition of the decoder on the encoded input;
- **Final linear layer**: a projection layer of the final decoder's output vector into a dimension corresponding to the cardinality of the vocabulary set. This vector is then processed through a **softmax** operation to transform the logits into a distributional range of probabilities summing to one. This vector's $argmax$ is the predicted token's index.

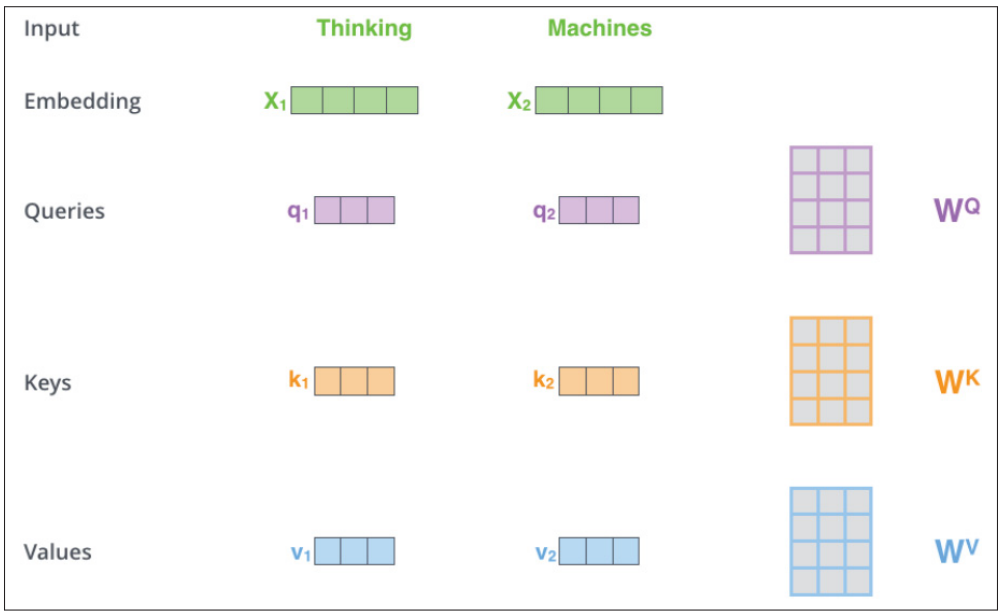


Figure 1.7 Query, Key and Value vectors
Taken from https://jalammar.github.io/illustrated-transformer/

1.1.3.1 More on self-attention

The self-attention scores are computed as follows:

1. Each token embedding is projected into three vectors, namely, **Query**, **Key**, and **Value** vectors, by respective multiplication with trainable matrices, $\mathbf{W^Q}$, $\mathbf{W^K}$ and $\mathbf{W^V}$ (Figure 1.7);
2. Query, Key and Value vectors are respectively aggregated into **Q**, **K** and **V** matrices for batched computation;
3. Query vectors in **Q** are compared with key vectors in **K** through matrix multiplication $\mathbf{QK}^\top$;

4. $\mathbf{QK}^\top$ is divided by the square root of d_k, the key vector's dimensionality: $\frac{\mathbf{QK}^\top}{\sqrt{d_k}}$. This operation stabilizes the gradient in the backpropagation pass by normalizing the influence of the vectors' dimensionality;
5. $softmax(\frac{\mathbf{QK}^\top}{\sqrt{d_k}})$: $Softmax$ probabilistically weighs the importance of each token in relation to the attended one;
6. $softmax(\frac{\mathbf{QK}^\top}{\sqrt{d_k}})\mathbf{V}$: $Softmax$ is used as weights in the weighted sum over the Value vectors.

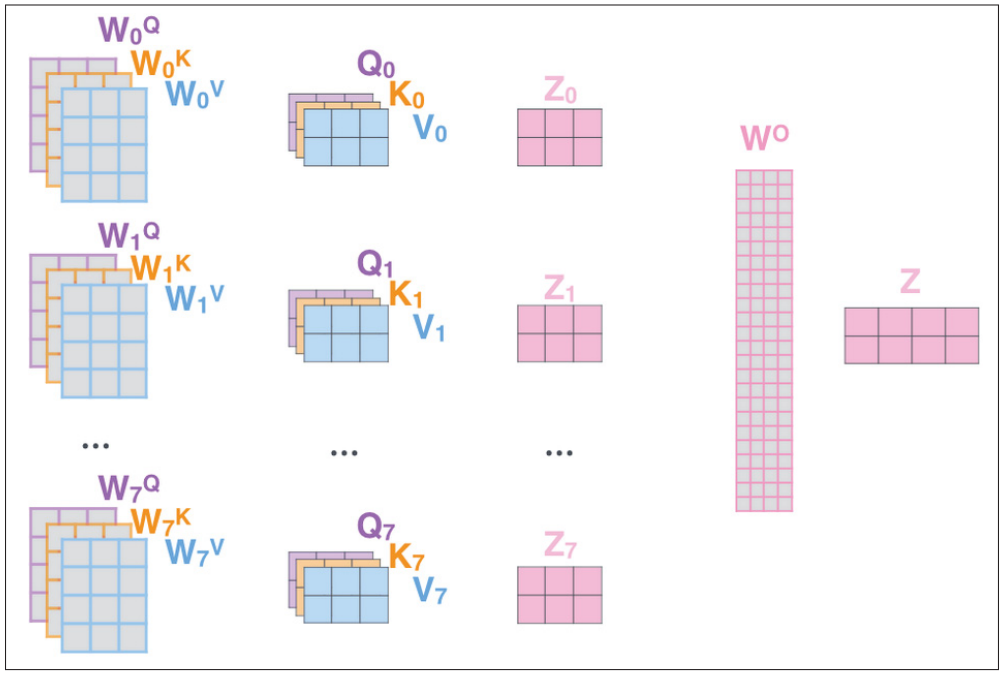


Figure 1.8 Multi-headed attention mechanism
Adapted from https://jalammar.github.io/illustrated-transformer/

The self-attention process is executed eight times – in the original Transformer architecture – in parallel, in what is referred to as the multi-headed attention mechanism (Figure 1.8). The eight attention heads ($\mathbf{Z_i}$ in the figure) are concatenated and projected with a trainable matrix $\mathbf{W^O}$ into a representation matrix ($\mathbf{Z}$), comprising a vector representation of each word, as expected by the final FFNN layer.

1.1.4 Bidirectional Encoder Representations for Transformers (BERT)

Devlin, Chang, Lee & Toutanova (2019) introduce a Transformer-encoder-based model called Bidirectional Encoder Representations from Transformers (BERT). In the original architecture, BERT is a stack of twelve Transformer encoder layers (Figure 1.9).

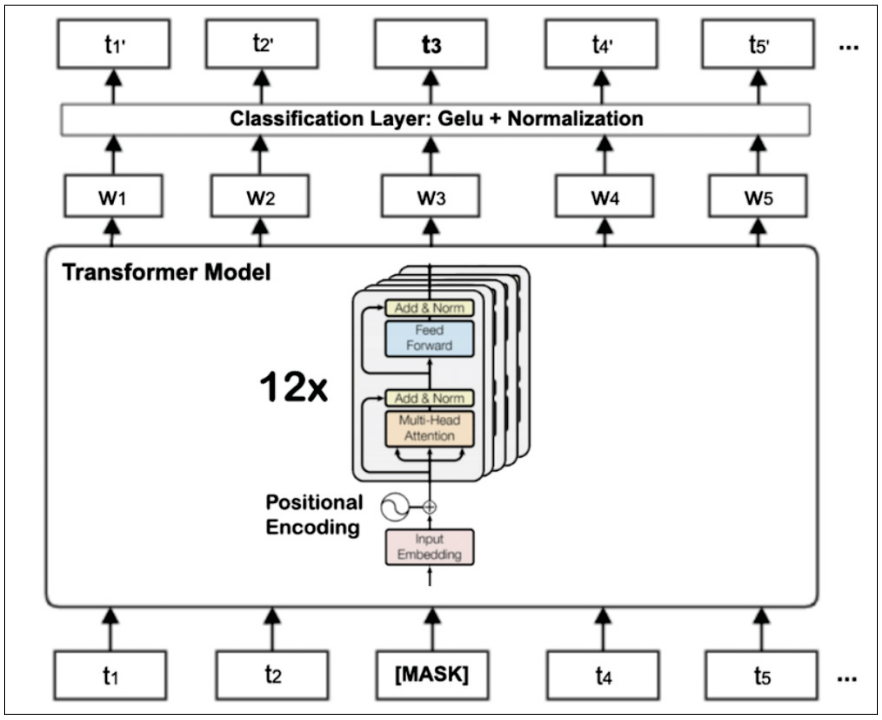


Figure 1.9 BERT architecture
Taken from Khalid *et al.* (2021)

While the original Transformer was designed to solve sequence transduction tasks, BERT is a Language Model for representation learning characterized by the statistical precedence relationship between words in a language. Representation learning is achieved by its pre-training on the Masked Language Modeling and Next Sentence Prediction tasks simultaneously.

BERT's pre-trained state introduces the Transfer Learning method into NLP; since BERT can be trained through self-supervision for representation learning, it can then be fine-tuned (further trained on relatively small datasets) for specific tasks such as **automatic summarization**.

The publishing of BERT motivated **new techniques in automatic summarization** as well as a **re-visiting of previous ones** for coupling them with the powerful encoding capabilities of BERT.

1.2 Automatic Text Summarization

Automatic text summarization is a conditional Natural Language Generation task. Nenkova & McKeown (2011) define it as generating a shorter version of a document while retaining its most important content.

The automatic summarization task can be laid on 2 axes:

- **Single-document or Multi-document**: the summary can be generated from an input consisting of a single or multiple documents. This constitutes a defining criterion for a summarization model;
- **Extractive or abstractive**: an extractive summary is formed from extracting and concatenating the most salient passages from the input document(s). An abstractive summary is more akin to human-redacted summaries in that it may rephrase, use synonyms, leverage general knowledge, etc.

1.2.1 TextRank: Bringing Order into Texts

TextRank (Mihalcea & Tarau, 2004) is a graph-based (Figure 1.10) text-ranking model based on Pagerank Page, Brin, Motwani & Winograd (1999).

The idea of a graph-based ranking model is to assign an importance score to nodes. Nodes can represent text units (e.g. words, tokens, sentences...etc.). A given node's score is increased by the connections linking to it, which can be of lexical or semantic natures.

The final iteration of the ranking algorithm assigns the true importance score to all text units. The highest-scoring text units are selected to form the output summary.

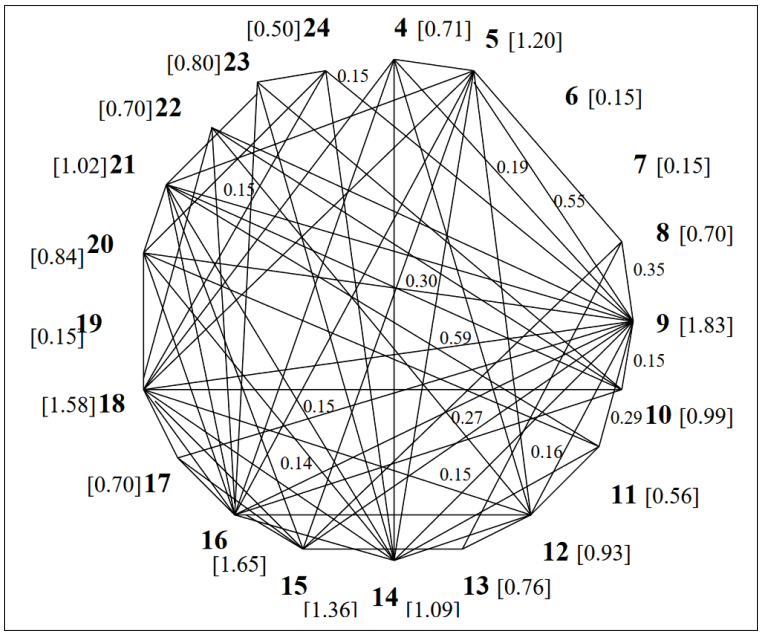


Figure 1.10 TextRank graph
Taken from Mihalcea & Tarau (2004)

1.2.2 LexRank: graph-based lexical centrality as salience in text summarization

Erkan & Radev (2004) propose LexRank, a multi-document extractive summarizer representing a text with a graph where each node encodes a sentence.

Nodes are connected with undirected weighted edges; weights represent the degree of similarity in meaning between a given pair of nodes (sentences). The similarity computation is based on the Term Frequency-Inverse Document Frequency method, or TF-IDF (Luhn (1957) and Sparck Jones (1972)).

LexRank differs from TextRank in that the latter computes weights as unit weights, whereas the former's use of degrees of similarity instead allows for more granular attribution of importance scores.

A uniform random walk is performed to compute the centrality of each node. The centrality of a node can be defined as the sum of all weight edges surrounding it. Therefore, a **node** with a

high centrality theoretically represents a **salient part of the text**. The most **central sentences** are then extracted to form the **output summary**.

1.2.3 Sentence Centrality Revisited for Unsupervised Summarization (PacSum)

The PacSum model, proposed by Zheng & Lapata (2019), is an algorithm similar to LexRank with two main differences:

- The edges are directed to represent a precedence relationship between sentences. The authors argue that encoding positional information into the model further enhances the quality of the generated summary;
- BERT replaces TF-IDF for encoding sentences and calculating pairwise node affinities with a similarity matrix.

1.2.4 Text Summarization with Pretrained Encoders (BertSum)

BertSum (Liu & Lapata, 2019) is a BERT-based model fine-tuned for the summarization task.

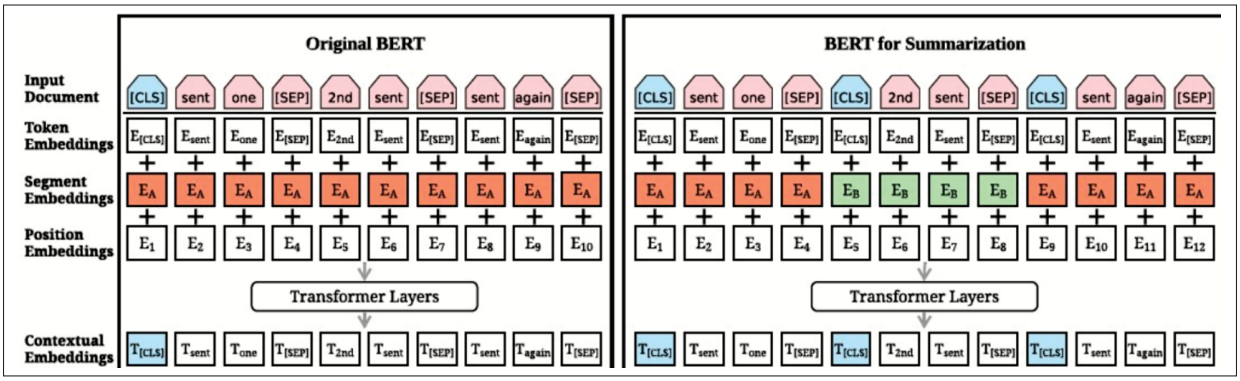


Figure 1.11 BertSum architecture
Taken from Liu & Lapata (2019)

Liu & Lapata (2019) provide a framework (Figure 1.11) for both extractive and abstractive summarization:

- The abstractive BertSum, *BertSumAbs*, is obtained from fine-tuning BertSum on *XSUM* (Narayan, Cohen & Lapata, 2018), an abstractive summarization dataset;

- The extractive BertSum, *BertSumExt*, is obtained from fine-tuning BertSum on *CNN/Daily-Mail* (Hermann *et al.*, 2015), an extractive summarization dataset.

1.3 Query-Focused Automatic Summarization (QFS)

Jones (1998) introduces automatic summarization's context factors (Figure 1.12) : input, output, and **purpose factors**. Respectively, the nature of the input text, that of the output summary, and its purpose.

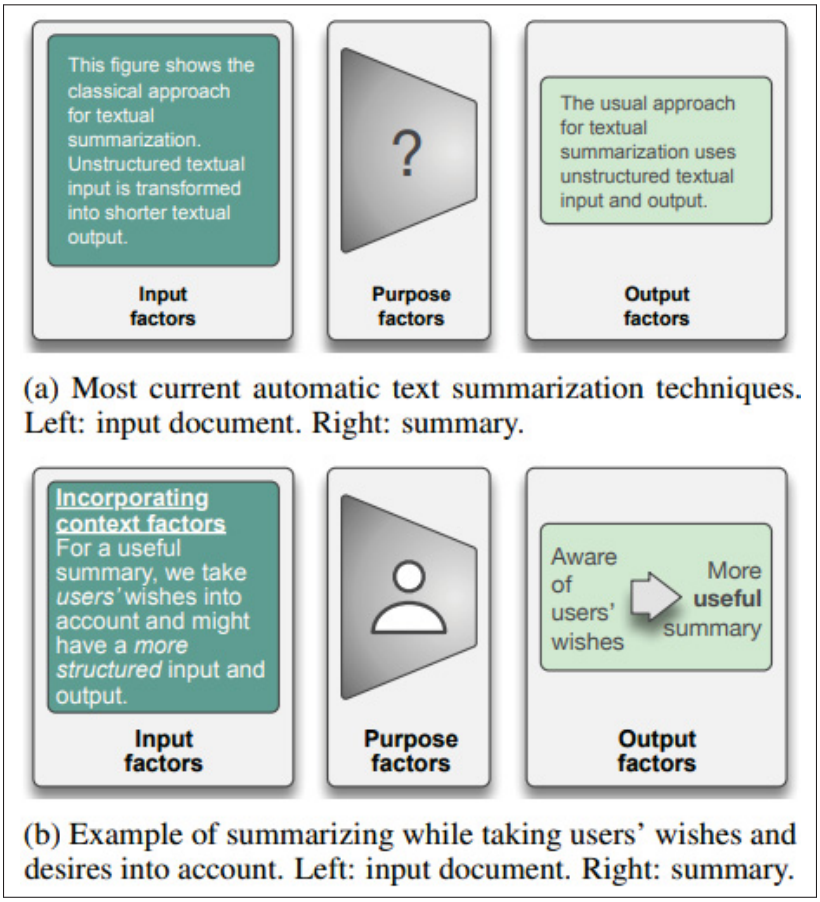


Figure 1.12 Context factors in automatic summarization
Taken from Ter Hoeve *et al.* (2022)

Ter Hoeve *et al.* (2022), who ground their work in that of Jones (1998), advocate for the usefulness of a summary with respect to the user's needs and report that the purpose factors receive the least attention from works in the field of automatic summarization, barring some of

its specializations which factor in the audience and the situation; among which is the task of Query-Focused Summarization (QFS).

The expected output of the QFS task is a summary addressing the input query based on the input document(s).

1.3.1 The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries

Maximal Marginal Relevance (Carbonell & Goldstein, 1998) is an MDQFES algorithm that conjointly considers diversity and query-relevance when selecting documents to retrieve salient passages from, as formulated by the following equation:

$$MMR = argmax_{D_i \in R \setminus S}[\lambda Sim_1(D_i, Q) - (1 - \lambda) max_{D_j \in S} Sim_2(D_i, D_j)] \tag{1.4}$$

D: Documents in the C collection — λ: Control parameter for accuracy vs diversity

Q: Query — **S**: Current result set

R: Relevant documents in C — $\boldsymbol{Sim_{1,2}}$: Similarity metrics determined by the user

Thus, MMR selects the document among the collection with the λ-controlled compromise between query relevance and diversity.

1.3.2 Biased LexRank: Passage Retrieval using Random Walks with Question-Based Priors

Otterbacher, Erkan & Radev (2009) extend LexRank by proposing the Biased LexRank algorithm. The random walk's originally uniform distribution is instead biased towards the similarity score of a given sentence with respect to the input query (Figure 1.13).

The generalized form of the LexRank equation can be written as:

$$LR(u) = \frac{d}{N} + (1 - d) \sum_{v \in adj[u]} \frac{w(v,u)}{\sum_{z \in adj[v]} w(v,z)} LR(v) \tag{1.5}$$

$\boldsymbol{LR(u)}$: LexRank value of sentence u — $\boldsymbol{N}$: Number of sentences (graph nodes)

$\boldsymbol{d}$: damping factor — $\boldsymbol{adj[u]}$: Set of u's neighboring node-sentences

$\boldsymbol{w(v,u)}$: Weight linking sentences v and u — $\boldsymbol{w(v,z)}$: Weight linking sentences v and z

Otterbacher *et al.* (2009) argue that an interesting interpretation of the LexRank value of a sentence can be understood in terms of the concept of a random walk, that is, the process of visiting the nodes (sentences) of the graph according to a specified transition probability distribution.

They then introduce a $b(u)$ term to bias the random walk while computing the LexRank score of a node (sentence):

$$LR(u) = d\frac{b(u)}{\sum_{z \in C} b(z)} + (1 - d) \sum_{v \in adj[u]} \frac{w(v,u)}{\sum_{z \in adj[v]} w(v,z)} LR(v) \tag{1.6}$$

*C denotes the set of all nodes in the graph

As such, the nodes' centrality additionally incorporates the query's information. Biased LexRank is a multi-document query-focused extractive summarizer.

1.3.3 Improving query focused summarization using look-ahead strategy

Badrinath, Venkatasubramaniyan & Veni Madhavan (2011) augment Biased LexRank by further biasing the random walk towards the query-relevance of the neighboring N-sentences instead of Biased LexRank's approach of only considering individual sentences.

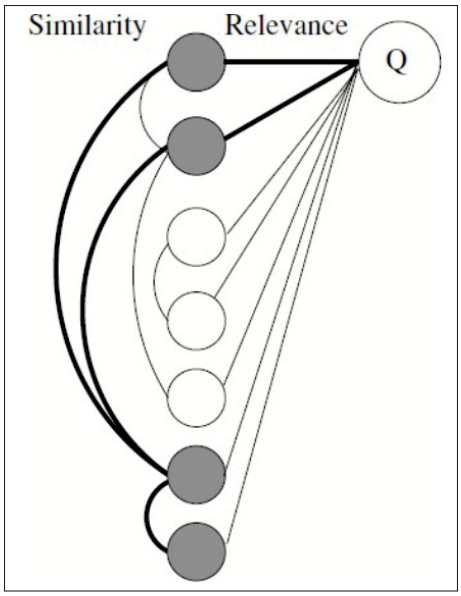


Figure 1.13 Similarity relevance
Taken from Otterbacher *et al.* (2009)

Badrinath *et al.* (2011) argue that the random walk surfer in Biased LexRank completely ignores the neighborhood information of the destination node and that their look-ahead approach uncovers nodes that are indirectly related to the query.

1.3.4 Diversity driven attention model for query-based abstractive summarization (DDA)

The Diversity-Driven Attention model (DDA) (Nema, Khapra, Laha & Ravindran, 2017) is motivated by the repetition problem in Attention-based decoders.

Nema *et al.* (2017) argue that the repetition problem occurs because two consecutive decoder states are likely to be similar; the DDA model (Figure 1.14) approaches this issue by imposing orthogonality (vectors dissimilarity) between the current and previous context vectors (numerical representations of the observed word's neighboring words). The level of orthogonality is gated by a γ parameter.

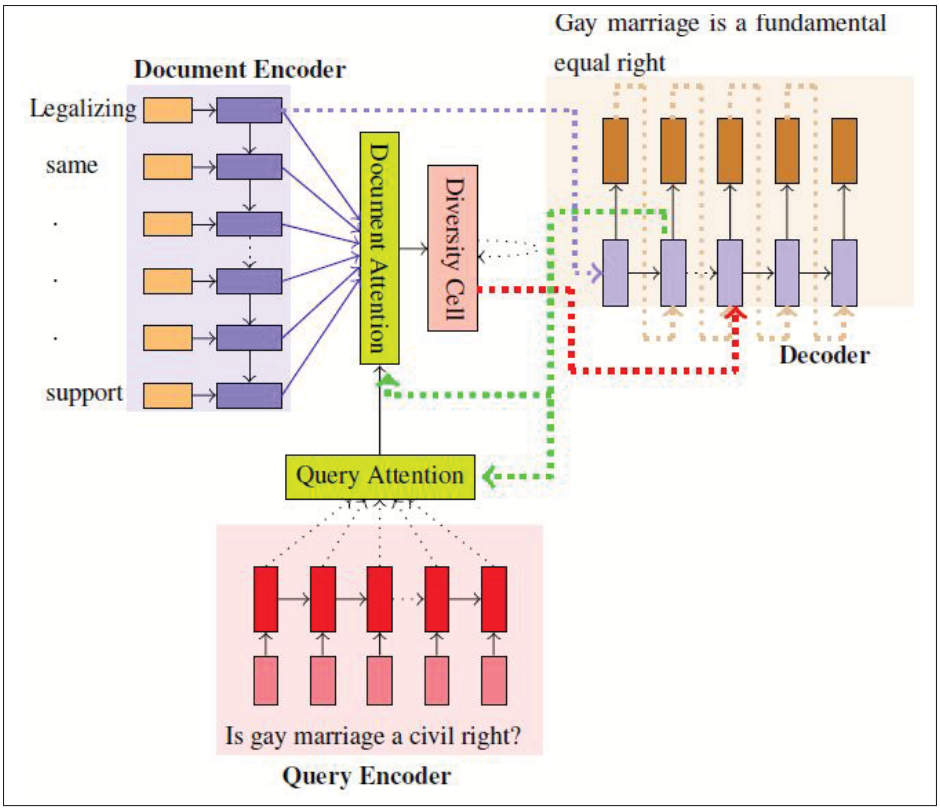


Figure 1.14 Diversity Driven Attention schema
Taken from Nema *et al.* (2017)

Nema *et al.* (2017) additionally introduce the *DebatePedia* dataset for Query-Focused Abstractive Summarization and augments it for training.

1.3.5 Query Focused Abstractive Summarization via Incorporating Query Relevance and Transfer Learning with Transformer Models (QR-BERTSUM-TL)

Laskar, Hoque & Huang (2020a) extend the BertSum model that was pre-trained on the XSum dataset (Narayan *et al.*, 2018) by 1) fine-tuning it on the DebatePedia dataset and 2) concatenating a query to the input document.

QR-BERTSUM-TL outperforms the DDA model without augmenting training data, and the $\text{BertSum}_{\text{XSUM}}$ model, which did not undergo fine-tuning. Figure 1.15 shows an output example from an input query and document, using QR-BERTSUM-TL.

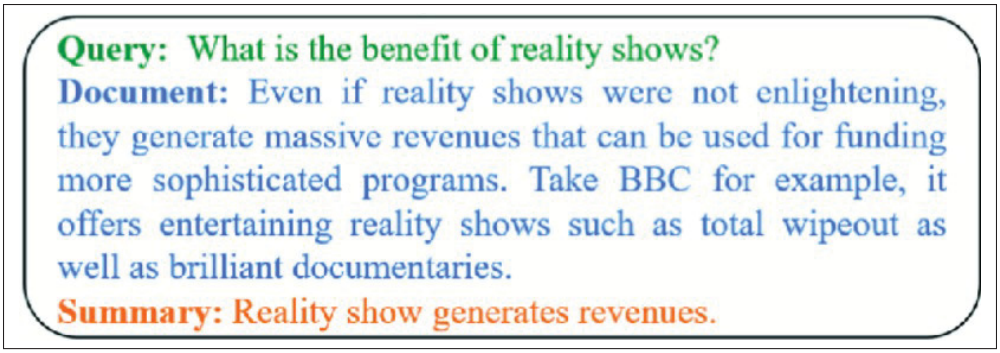


Figure 1.15 Query-Focused Summarization example
Taken from Laskar *et al.* (2020a)

1.3.6 Biased TextRank: Unsupervised Graph-Based Content Extraction

Kazemi, Pérez-Rosas & Mihalcea (2020)'s approach to bias TextRank's (Mihalcea & Tarau, 2004) summarization algorithm is reminiscent of that of Biased LexRank's (Otterbacher *et al.*, 2009) with respect to LexRank (Erkan & Radev, 2004).

TextRank computes the scores of nodes (text units) in the order produced by a random walk where each node has an equal chance of being visited. Nodes' scores are computed using neighboring nodes, which means that the scores of the nodes are sensitive to their visiting order.

Biased TextRank weights a node's visiting priority by its embedding proximity to the input bias. This ensures that the final top nodes will be bias-relevant based on their similarity to the bias. Kazemi *et al.* (2020) use cosine distance between the Sentence-BERT (Reimers & Gurevych, 2019) embeddings of the node (sentence as a text unit) and the input bias (e.g. query).

Biased TextRank in its summarization application can be an MDQFES model. It operates solely on text units (e.g. sentences) meaning that multiple documents can be concatenated into one for the graph representation.

1.3.7 WSL-DS: Weakly Supervised Learning with Distant Supervision for Query Focused Multi-Document Abstractive Summarization

To address the lack of labeled Query-Focused Summarization datasets, Laskar, Hoque & Huang (2020b) propose an approach for distant and weakly supervised learning (WSL-DS) to generate weak (artificially generated) reference summaries from gold reference summaries through a pre-trained, RoBERTa-based (Liu *et al.*, 2019), sentence-similarity model.

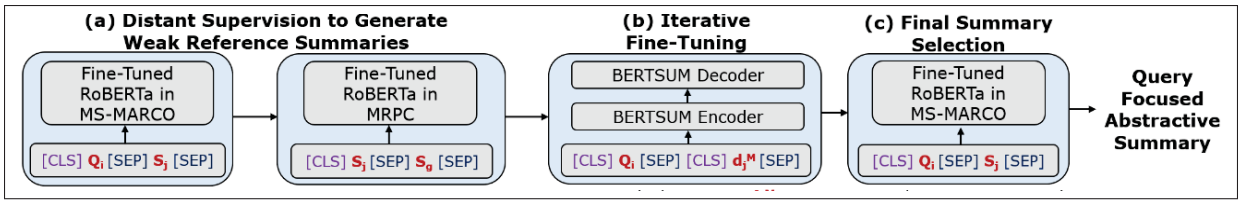


Figure 1.16 Process of the WSL-DS model for QFAS
Adapted from Laskar *et al.* (2020b)

Figure 1.16 shows a process where:

1. Fine-tuned RoBERTa models are used to generate weak reference summaries from the source documents' sentences;
2. The BertSum model is fine-tuned using the latter step's weak reference summaries;
3. A fine-tuned RoBERTa model is used for producing a query-focused abstractive summary.

1.3.8 Coarse-to-Fine Query Focused Multi-Document Summarization (QuerySum)

Xu & Lapata (2020) propose *QuerySum*, a coarse-to-fine approach (Figure 1.17) where the input document(s) is/are first decomposed into text passages, then are sequentially fed through three modules:

1. A relevance estimator: retrieves the top N sentences most relevant to the input query (if any);
2. An evidence estimator: trained through distant supervision on Question-Answering datasets and used to rerank the sentences output from the relevance estimator;

3. A centrality estimator: An extension of LexRank modified to incorporate the evidence estimator.

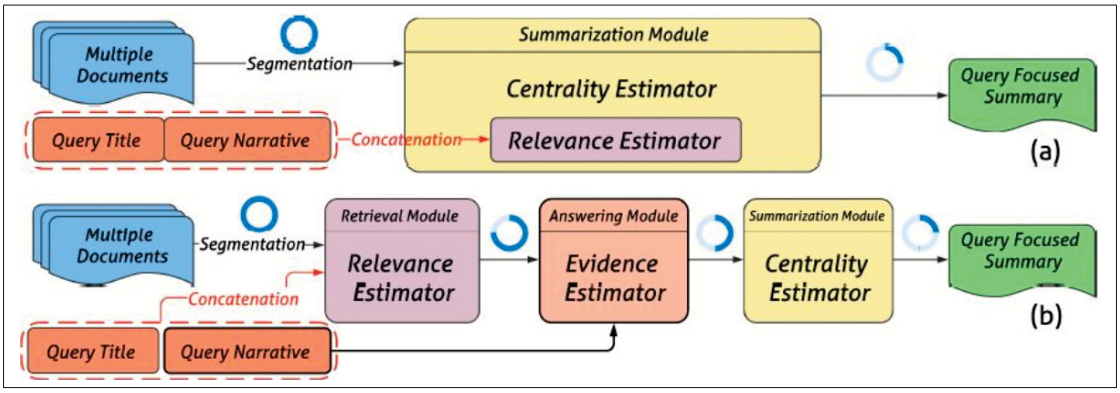


Figure 1.17 Architecture of the QuerySum model
Taken from Xu & Lapata (2020)

1.3.9 Generating Query Focused Summaries from Query-Free Resources

Xu & Lapata (2021) decompose the Multi-Document Query-Focused Summarization task into «(1) query modeling (i.e., finding supportive evidence within a set of documents for a query) and (2) conditional language modeling (i.e., summary generation) » (page 6096).

Xu & Lapata (2021) claim that query modeling assumes that a generic summary always contains information relevant to potential or latent queries.

The observed latent queries are reverse-engineered from the summaries via what the authors call a Unified Masked Representation (UMR) (Figure 1.18), which effectively renders the summaries into proxy queries for training.

The authors further assume that answers to these queries can be found in sentences from the document collection that score highly on ROUGE (Lin, 2004) with respect to the queries.

Their model produces queries in UMR, then ranks the input sentences by relevance. These are fed to a conditional language model to output query-focused abstractive summaries (Figure 1.19).

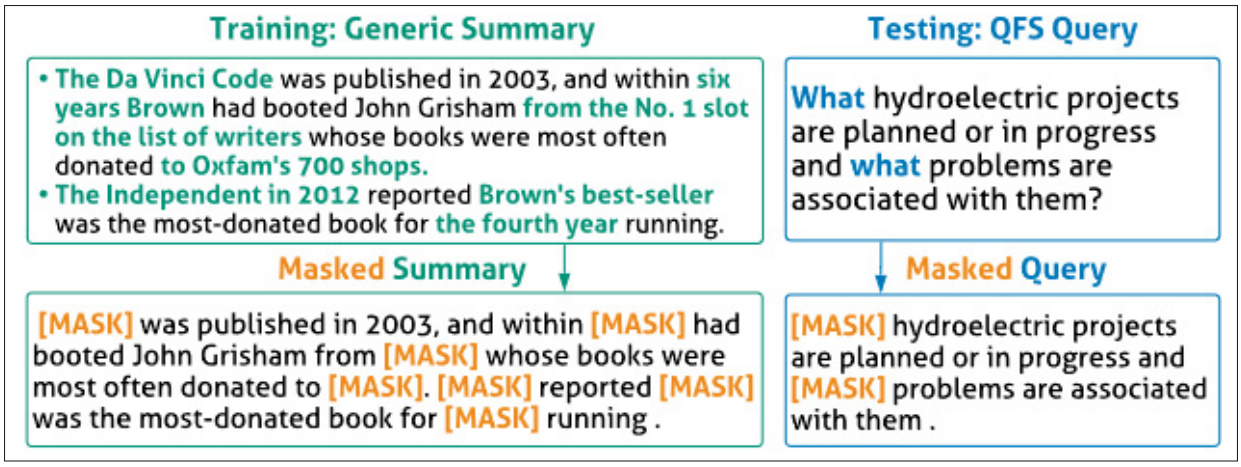


Figure 1.18 Unified Mask Representation
Adapted from Xu & Lapata (2021)

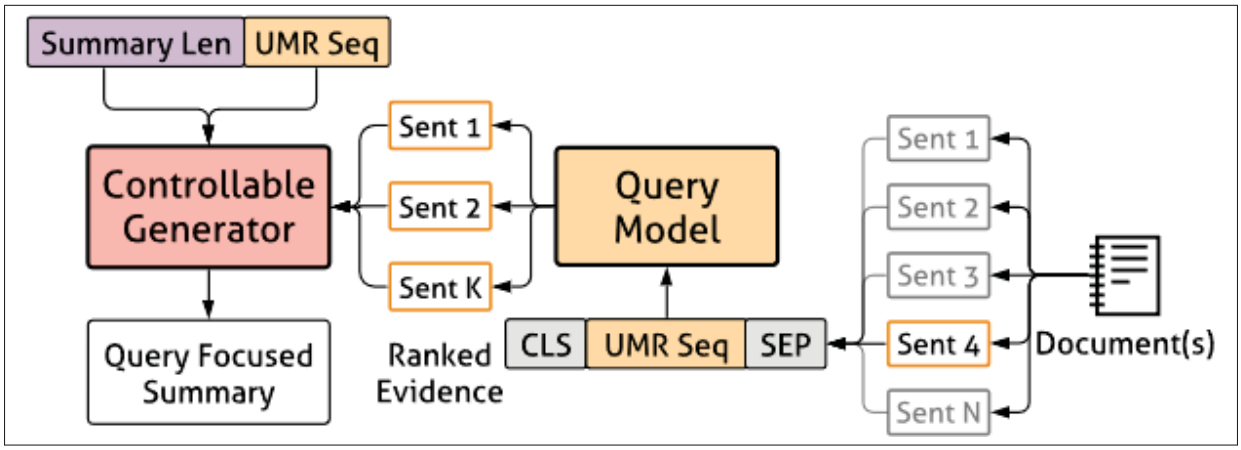


Figure 1.19 Framework of Query Focused Summarization
Adapted from Xu & Lapata (2021)

1.3.10 Text Summarization with Latent Queries

Xu & Lapata (2022)'s generative module, MARGE (Figure 1.20), is based on a dual view of the source document: a query-agnostic view joined with a query-focused view. The query module is optional, and not utilizing it falls back to generic summarization.

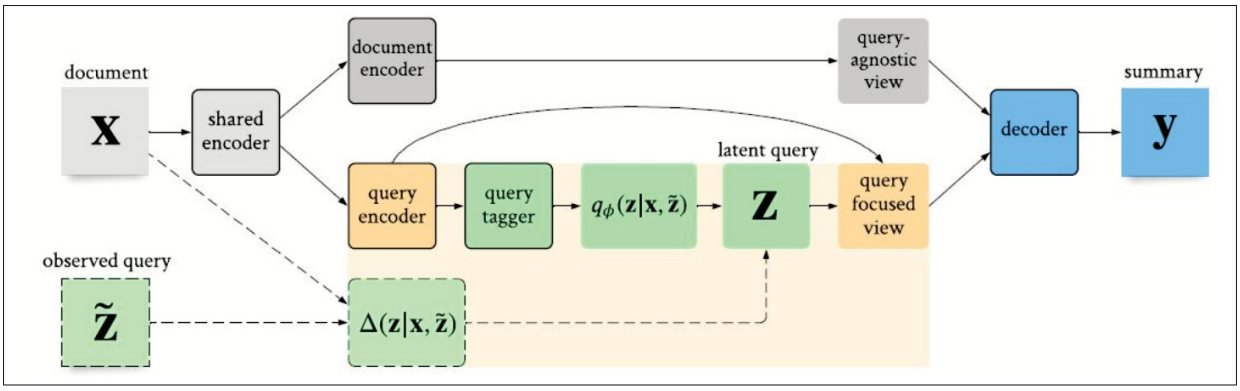


Figure 1.20 Architecture of the MARGE model
Taken from Xu & Lapata (2022)

Latent queries are encoded from the source document. A latent query is a binary variable with a probability distribution indicating the belief that a given token (text unit) from the source document is query-relevant. This means that the training set can be query-free.

1.3.11 Heterogeneous GNN for Query-focused [Extractive] Summarization

Ya, Liu, Cao & Guo (2021) propose a Graph Neural Network (GNN) composed of node representations (embeddings) of the input query and document-sentences. Query-sentence relevance is based on their common words. The semantic gap between query and sentence embeddings is bridged by the Mutual Information Maximization formula (Yeh & Chen, 2019). The authors adopt the sentence-classification approach on the QFES task, where each sentence can either belong to the summary or not. Node representations are updated to fit the classification task. Their model estimates the relevance probability for each sentence and outputs the top-ranking ones.

Figure 1.21 shows that the model architecture is thrice decomposed:

1. **Heterogeneous graph encoder**: initializes embeddings for heterogeneous nodes (document, sentence, word, and query);
2. **Graph neural layer**: a Graph Attention Network updates node representations through iterative message-passing;

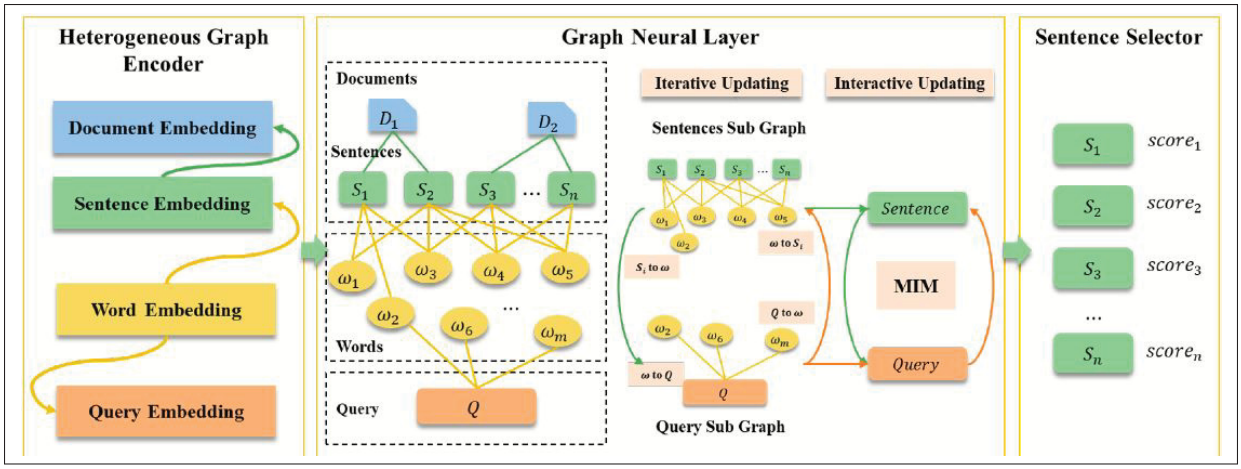


Figure 1.21 Heterogeneous GNN for QFES
Taken from Ya *et al.* (2021)

3. **Sentence selector**: Sentence-nodes are extracted to predict binary classification scores.

1.4 Query Expansion (QE)

Query expansion (QE) is the task of augmenting a query to cover the user's intent better and bridge the lexical and semantic gap between the query's formulation and the target documents' style. QE is usually ancillary to a downstream task such as information retrieval (e.g., search engines), question-answering, or query-focused summarization.

1.4.1 A Two-Stage Masked LM Method for Term Set Expansion

Kushilevitz, Markovitch & Goldberg (2020) propose two methods for the NLP task of Term Set Expansion (TSE), which consists of expanding members of a semantic class from a small set of terms, and can thus be utilized for query expansion:

1. Query-terms are masked in sentences in which they occur, marked as indicative patterns. A BERT-based Masked Language Model (MLM) predicts the masks for these sentences. The elected sentences are those for which the best prediction matched the masked query term. The next best predictions are selected to expand the query;

2. Semantically similar patterns are retrieved from the corpus. The words that would fill the mask -as predicted by the MLM- in these similar patterns are selected to expand the query.

1.4.2 Using Query Expansion in Manifold Ranking for Query-Oriented Multi-Document Summarization

Quanye, Rui & Jianying (2021) augment *TF-IDF*-encoded (Luhn (1957) and Sparck Jones (1972)) queries by combining four different transformations involving the query itself and/or the input documents. Expansions are based on semantic similarities with *Wordnet* (Miller, Beckwith, Fellbaum, Gross & Miller, 1990) synonyms, on the average and variance of *Term-Frequence Inverse Sentence Frequency* values of words, and on the TextRank (Mihalcea & Tarau, 2004) algorithm to extract query-relevant words from the documents and select the top ones with manifold ranking.

1.4.3 BERT-QE: Contextualized Query Expansion for Document Re-ranking

BERT-QE (Zheng *et al.*, 2020) takes a ranked list of documents as input (e.g., from an unsupervised ranking model such as TextRank) and outputs a re-ranked list based on the expanded query (Figure 1.22).

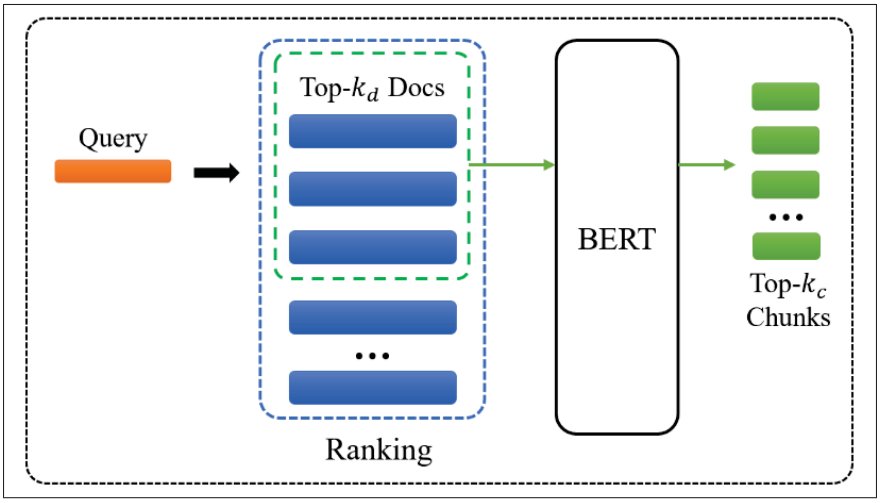


Figure 1.22 BERT-QE process
Taken from Zheng *et al.* (2020)

BERT-QE's objective is achieved in 3 phases:

1. Re-ranking with a BERT model, fine-tuned on MS MARCO (Bajaj *et al.*, 2018) to extract the top-ranked documents;
2. Chunking into fixed-length texts to be individually evaluated;
3. Assessing the relevance of a given document by scoring it based on the selected chunks and the original query.

1.5 Automatic summarization evaluation metric

The de-facto standard metric for automatic evaluation of text summarization models was proposed by (Lin, 2004) and is called *ROUGE*, or Recall-Oriented Understudy for Gist Evaluation.

ROUGE is a measure of n-gram overlaps between the generated and reference summaries. An n-gram, in the context of ROUGE, is a contiguous sequence of n-words. ROUGE's commonly used variations are ROUGE-1, ROUGE-2, and ROUGE-L, respectively, unigram-overlap-based comparison, bi-gram-overlap-based comparison, and longest common subsequence-overlap-based comparison.

It is noteworthy that there is no unanimous accord on which variant correlates best with human judgment:

- Lin (2004) report *ROUGE-2, ROUGE-L, ROUGE-W*, and *ROUGE-S* as the best for single-document summarization tasks, *ROUGE-1, ROUGE-L, ROUGE-W, ROUGE-SU4, and ROUGE-SU9* for very short summaries, *ROUGE-1, ROUGE-2, ROUGE-S4, ROUGE-S9, ROUGE-SU4, and ROUGE-SU9* for multi-document summarization, and that correlations to human judgments were generally increased by stopwords removal and using multiple references;
- Owczarzak, Conroy, Dang & Nenkova (2012) report the metric closest to human judgment as *ROUGE-2-Recall* with stemming and retaining of stopwords;
- Rankel, Conroy, Dang & Nenkova (2013) report *ROUGE-3* and *ROUGE-4*, which are rarely reported, as among the most accurate human-judgment-wise;

- Graham (2015) elects *ROUGE-2-Precision* with stemming and stop-words removed as the metric most faithful to human judgment.

$$ROUGE = \frac{count_{match}(gram_n)}{count(gram_n)} \tag{1.7}$$

In employing the above equation, given the -non-sensical- sentence "the hello a cat dog fox jumps" and the reference text "the fox jumps", there would be three overlap hits ("the", "fox", and "jumps") among the seven tokens, and so a rounded 43% ROUGE-1 precision score.

Despite incorporating Wordnet (Miller *et al.*, 1990) for synonymous hits in ROUGE implementations, this example highlights the semantic agnosticism of ROUGE and the motivation for works in richer automatic evaluation metrics.

No other metric has yet been established as the new standard for the automatic evaluation of summaries, as the current trend in works prioritizes enabling comparison with previous works.

1.5.1 METEOR

Banerjee & Lavie (2005) propose *METEOR*, a method for evaluating generated -or extracted- text with respect to reference texts based on exact word matches in terms of precision and recall scores. Matching of synonyms, stemmed words, and paraphrases are also factored in.

Denkowski & Lavie (2014) propose METEOR 1.5, which assigns different weights to functions, content words, and distinct matching categories.

Guo & Hu (2019) propose METEOR++ 2.0, which extends METEOR by 1) integrating an external paraphrasing resource and 2) using BERT (Devlin *et al.*, 2019) embeddings.

1.5.2 Evaluating the Factual Consistency of Abstractive Text Summarization

Kryściński, McCann, Xiong & Socher (2020) propose *FactCC*, a BERT-based weakly supervised model for assessing factuality in model-generated summaries. Training data comprises textual transformations (e.g. paraphrasing through back-translation, noise injection, and sentence negation) on spans from the source documents. FactCC functions by:

1. Predicting the factual consistency of summary sentences;
2. Extracting source-document spans to support the prediction;
3. Extracting inconsistent spans from summary sentences deemed inconsistent.

The authors define a factually consistent summary as a summary containing only declarations logically implied by the source document.

1.5.3 FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization

Durmus, He & Diab (2020)'s model, *FEQA* (Figure 1.23), utilizes a Question-Answering (QA) model to evaluate the factual faithfulness in abstractive summaries. This model is unsupervised and does not require reference summaries. A summary is considered unfaithful if it contains errors of contradiction or hallucination. The authors also provide a method for evaluating the degree of abstraction of a summary.

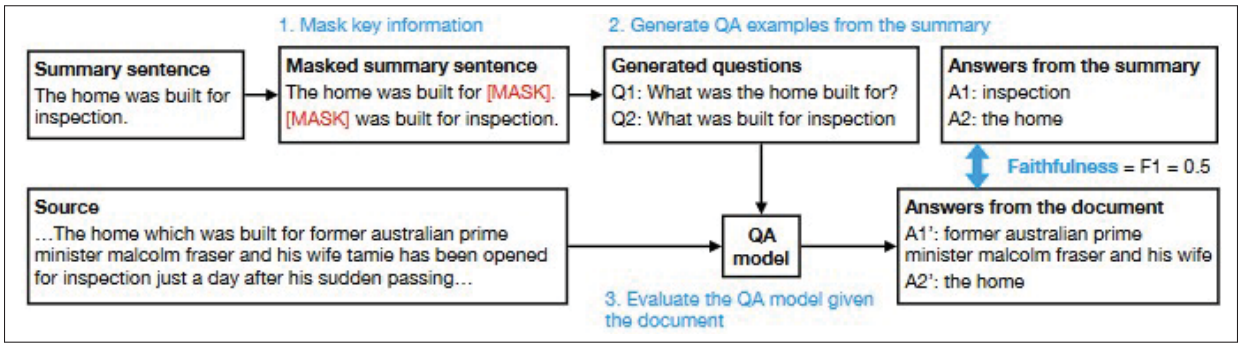


Figure 1.23 Overview of FEQA's process
Taken from Durmus *et al.* (2020).

The following steps describe FEQA's architecture:

1. The summary's entities are detected through Named Entity Recognition (NER) and masked. A BART model (Lewis *et al.*, 2019) fine-tuned on the QA2D dataset (Demszky, Guu & Liang, 2018) is used to generate questions, e.g., "<MASK> was built" → "What was built?". The answer is the item that was originally masked;
2. The generated questions and the source document are provided to a QA model, i.e., BERT (Devlin *et al.*, 2019) fine-tuned on the SQuAD dataset (Rajpurkar, Zhang, Lopyrev & Liang (2016) and Rajpurkar, Jia & Liang (2018)) to generate answers. These are then compared with the masked ground truths to estimate faithfulness.

1.5.4 Fact-based Content Weighting for Evaluating Abstractive Summarisation

Xu, Dušek, Li, Rieser & Ioannis Konstas (2020) propose *Corr-F/A*, an unsupervised evaluation model where facts/arguments from the generated summary are extracted and compared against facts/arguments from the input document. The latter are weighted by correspondence with facts/arguments from the reference summary. Such weights are calculated based on the following:

- Semantic similarity between the vector embeddings of the input document's facts and the reference summary's facts;
- Distance within a meaning representation tree based on Semantic Role Labeling (Palmer, Gildea & Kingsbury, 2005).

Thus, Corr-F/A indicates how much the output summary's facts/arguments agree with those that the reference summary deems important in the input document.

1.5.5 SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization

Gao, Zhao & Eger (2020) propose *SUPERT* (Figure 1.24), an unsupervised model to evaluate model-generated summaries by comparing them with pseudo-reference summaries. Their Salient

Sentence Extractor generates the latter, and the comparison consists of semantic similarity computation between SBERT embeddings (Reimers & Gurevych, 2019).

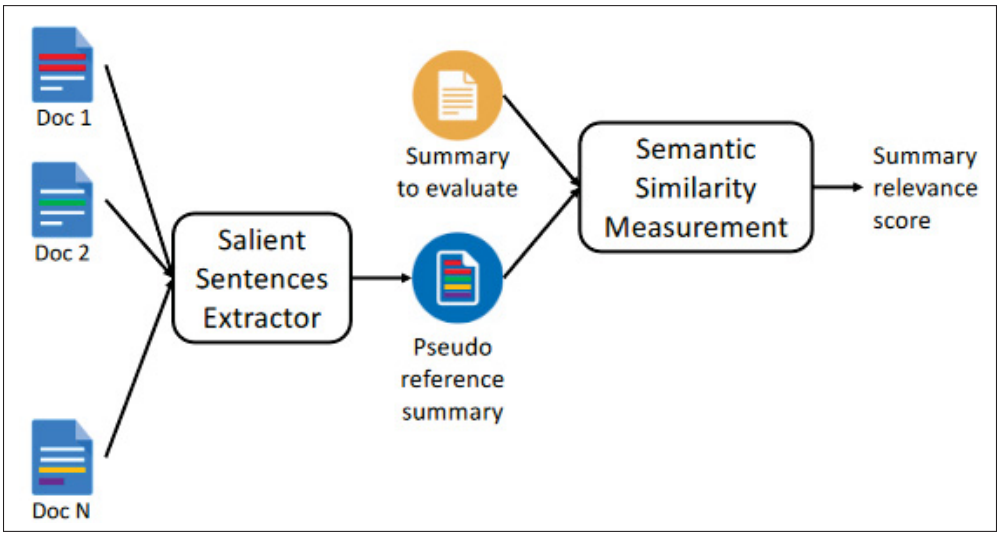


Figure 1.24 Workflow of SUPERT
Taken from Gao *et al.* (2020).

1.5.6 Factual Error Correction for Abstractive Summarization Models

Cao, Dong, Wu & Chi Kit Cheung (2020) propose a BART-based (Lewis *et al.*, 2019) corrector model trained on artificial examples generated from the reference summaries. The corrector model could be used to evaluate the factual consistency of generated summaries, assuming that a generated summary is inconsistent if the corrector had to edit it.

Cao *et al.* (2020) incorporate entity transformations from FactCC (Kryściński *et al.*, 2020) and apply their future work suggestions by adding transformations on numbers, dates, and pronouns to train the corrector model on these error types.

The objective of the corrector model is to produce a corrected summary based on the source document and the transformed reference summary.

1.5.7 BERTScore: Evaluating Text Generation with BERT

Zhang, Kishore, Wu, Weinberger & Artzi (2020) propose *BERTscore* (Figure 1.25), a BERT-based model for automatic evaluation of text generation. BERTscore computes cosine similarity between contextual embeddings of the generated tokens and the reference ones. The similarity is optionally weighted with inverse document frequency scores.

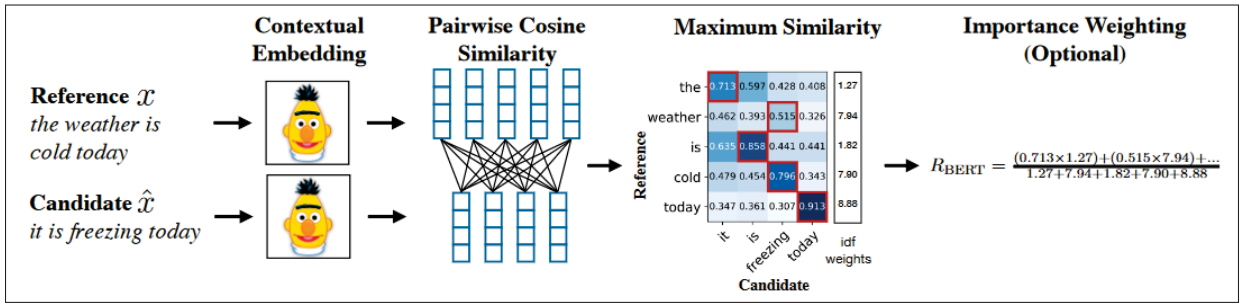


Figure 1.25 BERTSCORE process
Taken from Zhang *et al.* (2020)

1.5.8 BLEURT: Learning Robust Metrics for Text Generation

Sellam, Das & Parikh (2020) propose *BLEURT*, a BERT-based model for automatically evaluating text generation tasks. The aim is to predict human ratings on the correspondence between the generated tokens and the reference ones. BLEURT is pre-trained on standard sentence-pair datasets. Contrary to the latter, the input data might not be cleaned and structured at the inference time of an arbitrary Natural Language Generation (NLG) task. To remedy this issue, the authors inject noisy data into their pre-training by utilizing techniques such as mask-filling, back-translation, and random dropping of words.

BLEURT is pre-trained on a variety of NLG task signals:

- Automatic metrics: BLEU (Papineni, Roukos, Ward & Zhu, 2001), ROUGE (Lin, 2004) and BERTscore (Zhang *et al.*, 2020);
- Backtranslation likelihood;
- Textual entailment;

- Backtranslation flag.

1.6 Taxonomy of errors in automatic summarization

Table 1.1 Error types in automatic summarization. Bold represents extracted text.

Error types	**Description**
Erroneous anaphoric references	" **[De Long and his crew sailed with the ship Jeannette through the Bearing sea 1879.]** [...] Fridtjof Nansen immediately got the idea to test the theory of an open sea filled with drift ice. [**He let build a powerful ship strong enough to drift unharmed with the thick pack ice for a long time.**] " (Kaspersson, Smith, Danielsson & Jönsson, 2012) The preceding antecedent was not extracted; "He" refers to "Fridtjof Nansen", but since this part was not extracted, the anaphoric pronominal reference is erroneously attributed to "De Long".
Broken anaphoric references	" The free man had, however, official duty. [**Such official duty was the guesting, the obligation to receive and support the king and his escort when they traveled**]". (Kaspersson *et al.*, 2012) The pronoun "such" in the extracted text has no antecedent. The anaphoric reference is therefore broken.
Cohesion and context	Extracted phrases that lack the necessary cohesion or context to understand their meaning. Smith, Danielsson & Jönsson (2012) measure cohesion by the number of co-references in the text and how intact it is after summarization.
Grammar	Whether the set of phrases follows the rules of the language.
Focus	The extracted text should only contain information relevant to the rest of the summary. (Over, Dang & Harman, 2007) The query relevance should also be considered in the case of query-focused summarization.
Structure and coherence	Over *et al.* (2007) state that the summary should not be a pile of disordered information but rather a sentence-by-sentence construction towards a coherent set of information on the subject.
Redundancy	Repetition of sentences with similar meanings. Especially occurrent in multi-document summarization.
Temporal & spatial relationships	When/Where an event took place. More prominent in multi-document summarization. (Kaspersson *et al.*, 2012)

Table 1.1 presents the most prominent error types in automatic summarization found in literature, paired with concrete examples or detailed explanations for each.

1.6.1 DUC in Context

Over *et al.* (2007) study the Document Understanding Conference (DUC) dataset in the context of automatic summarization. They report and utilize DUC 2005 and 2006's set of linguistic quality in their study. This set numbers five error types, among which are grammaticality, redundancy, focus and structure & coherence (see Table 1.1 for details) as well as referential clarity, which is analogous to anaphoric references.

1.6.2 This also affects the context - Errors in extraction-based summaries

Kaspersson *et al.* (2012) attempted to address several summarization error types by varying the degree of summarization (summary length). They found that this approach mitigates some error types and exacerbates others. They report absent cohesion, absent context, and broken anaphoric references as the most common error types.

The authors observe that some error types are directly proportional to the degree of summarization, while U-shaped or cut-off linear relations describe others.

1.6.3 Cohesion in Automatically Created Summaries

Smith *et al.* (2012) report results of cohesion studies in extractive summaries. They measure it by the number of co-references in the text and its intactness in the summary. They conclude that taking previous sentences in extraction slightly improves cohesion but does not significantly improve quality, despite the assumed tradeoff between the amount of information included in the summary and its cohesion.

CHAPTER 2

OBJECTIVES

Chapter 1 explored the literary landscape of automatic summarization models, starting from text-encoding models, leading to query-focused summarization and query-expansion models, then ending with how they can be evaluated.

With the awareness of the state-of-the-art permitting assessment of reasonable goals and potential areas of improvement, this chapter presents the main objective of this work and its sub-objectives.

2.1 Explicative Sentiment Summarization

Given a multitude of documents, a sentiment of interest, and a query regarding the targeted entities (e.g. a specific product, date, or location), the main objective is to provide an informative summary explaining the cause(s) of the queried sentiment. This requirement describes a constrained **Query-Focused Summarization (QFS)** task, which we term as **Explicative Sentiment Summarization (ESS)**.

2.1.1 Qualitative study of prominent error types in output summaries

Section 1.6 presented a taxonomy of errors in automatic summarization. Following this, we perform a qualitative study of existing models' output summaries to identify the most frequent error types (1.1) and thus prioritize the problems to solve in Query-Focused Summarization, and by extension, our purpose task of Explicative Sentiment Summarization.

2.1.2 Propose and evaluate solutions at the QFS level

Following findings in 2.1.1, we approach the identified prominent problems with solution proposals and evaluations at the scope of Query-Focused Summarization.

2.1.3 Propose and evaluate solutions at the ESS level

Following findings in 2.1.1 and 2.1.2, we propose and evaluate solutions at the specialized scope of Explicative Sentiment Summarization.

2.2 Quantitative study of MDQFES models

The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin (2004)) metric will be used as it is the de-facto standard in automatic summarization. See section 4.2 in chapter 6.

CHAPTER 3

METHODOLOGY

Chapter 2 presented this work's main objective and its transversal and sub-objectives. This chapter reports the methodologies involved in achieving the declared goals.

3.1 Quantitative study of MDQFES models

ROUGE uses various strategies to quantify n-gram overlap between the output text and its reference(s). We heuristically elect the F1 score of **ROUGE-SU4** as the most relevant ROUGE variant for our problem space. We also report F1 scores of the **ROUGE-1**, **ROUGE-2** and **ROUGE-L** variants. Refer to section 4.2, "Evaluation Metric for Automatic Summarization," in chapter 6 for the justification of these choices.

Using the aforementioned ROUGE variants, we evaluate MDQFES models on the Document Understanding Conference (DUC) datasets (from 2005, 2006, and 2007), and on our proprietary Explicative Sentiment Summarization (ESS) dataset, of which only metadata is disclosable.

The ESS dataset spans 950 ESS units, each containing:

- The name of the targeted entity;
- The sentiment of interest;
- 1 to 576 documents with a mean of 17 and variance of 38, with each document spanning 2 to 771 sentences with a mean of 24 and variance of 36;
- A single-sentence abstractive reference summary explaining the sentiment.

The DUC datasets cover a wide range of topics such as politics, science & technology, health, sports & entertainment, etc. They are a historically prominent standard for evaluating Multi-Document Query-Focused Summarization models. Table 3.1 presents an overview of the metadata concerning the used DUC datasets.

For simplification purposes, in our experiments, we concatenate the DUC2005, DUC2006, and DUC2007 datasets as one and hereafter refer to the group as DUC.

Table 3.1 DUC 2005-2007 statistics

DUC	**2005**	**2006**	**2007**
Clusters	50	59	30
Sentences	45,931	34,560	24,282
Words limit	250	250	250
Avg. queries/cluster	2.18	2.10	1.62
Avg. query length	3.55	12.53	15.11

3.2 Qualitative study of prominent error types in system summaries

To identify the error types (table 1.1) with the most negative impact on summary quality, we conduct a qualitative analysis on the worst output summaries in terms of ROUGE-SU4-F1 scores, as determined in section 3.1. The studied summaries are from the MMR, QuerySum, and Biased TextRank models for MDQFES, executed on the DUC and ESS datasets.

Appendix I presents examples of the methodology involved in the qualitative study.

3.3 Propose and evaluate solutions at the QFS level

In section 3.2, we identified "focus" as the most prominent error type. Thus, following sub-objective 2.1.1, the problem with the highest yield upon resolution is to be considered the problem. The focus error type, re-iterated from table 1.1, pertains to extractive summaries whose text passages lack intra-relevance and query-relevance in the case of Query-Focused Summarization.

The query-focus of text units is typically determined by a similarity computation on their embeddings and sentence encodings in the case of QFS. The underlying assumption is that the encodings are correctly distributed in semantic space regardless of the linguistic domain of the s ource (query) and the target (input documents). This state of affairs mainly involves the sentence encoder in use and the formulation of the input query. Thus, we deem these QFS components most likely to impact the query-focus of a summary and address our methodologies for approaching them in the following subsections.

3.3.1 Sentence Encoders

Using Biased TextRank, we fix the input query and vary sentence encoders on the test split of the ESS dataset. In addition to the SBERT-based (Reimers & Gurevych, 2019) sentence encoders, we test asymmetric semantic search encoders[1]. The latter are pre-trained on Question Answering datasets, which are theoretically more appropriate for the length asymmetry involved in retrieving long text passages with a short query. Section 4.3 reports a sample of the best-performing experimental results.

3.3.2 Query Formulation

See sections 3.1, 3.2, and 3.3 in chapter 6.

In particular, section 3.2 from chapter 6 introduces Multi-Bias TextRank (MBTR), a Compound Bias-Focused Summarization model (chapter 6, 3.1) which relies on the reduction of biases from multiple input queries. To determine the best folding operation for combining the biases, we perform experiments on the following reduction strategies, paired with their motive intuition:

1. **Summation:** amplification of the relevance-score of a desired sentence that might not have been effectively addressed by a single bias. Conversely, a low score denotes more confidence in the rejection of a sentence, given the implication that none of the biases were relevant to it;
2. **Max:** optimistic selection of the highest bias to nullify the contribution of the lesser relevant biases. A low score reinforces rejection confidence, akin to the summation effect;
3. **Mean:** uniform scaling of their summed contributions;
4. **Median:** lesser sensitivity to outlier biases;
5. **Conjoint probabilities:** framing of relevance-scores as independent salience likelihoods, thus enabling the conjunction of their influences;

[1] https://www.sbert.net/examples/applications/semantic-search/#symmetric-vs-asymmetric-semantic-search

6. **Negative variance:** favoring the sentence with the least volatile scores across biases, i.e. the most accord on relevance.

3.4 Propose and evaluate solutions at the ESS level

For sentiment explanation, we can disregard open-domain queries and specialize the QFS task for biases and queries that align with this objective. Additionally, we can leverage the prior knowledge of queries in a sentiment explanation setting. We introduce the task of **Explicative Sentiment Summarization (ESS)** in section 6.4.4.

CHAPTER 4

EXPERIMENTAL RESULTS

Chapter 3 presented the experimental methodologies whose results we present in this chapter.

4.1 Experimental protocol

Section 4.2 reports the qualitative study of DUC summaries across three summarizers: Maximal Marginal Relevance (MMR, implemented locally), QuerySum[2] and BiasedTextRank (BTR, implemented locally). For these experiments, hyperparameters for all three were preserved as recommended by their authors. This choice is justified by the proof-of-concept nature of this work. Hyperparameter optimization is relegated to a potential production phase.

Section 4.3 varies node representations in BTR with multiple sentence encoders, evaluating each with ROUGE scores using the pythonrouge[3] implementation.

Section 4.4 reports ROUGE scores for MultiBiasTextRank (MBTR) with query expansion (with detailed protocol in section 6.5.3 and in the footnotes of table 6.1). Section 4.4 also varies implementations of the reduction operator in 6.3, as reported in its table 4.2. The α and β choices are justified in section 6.6 and in table 6.1.

Section 4.5 refers to the manuscript article's sections 6.5.3, 6.5.4, and to its table 6.1 for the protocols leading to ROUGE scores of sentiment-aware MBTR in the context of the Explicative Sentiment Summarization task.

4.2 Qualitative study of prominent error types in system summaries

The x-axis in Figure 4.1 comprises error types regarding pronominal references, nominal references, query-focus, cohesion & context, and redundancy. The y-axis presents the total count for each error type on the DUC dataset across three QFMDES models: MMR, QuerySum, and

[2] https://github.com/yumoxu/querysum

[3] https://github.com/tagucci/pythonrouge

BTR. Results show a consistent **prominence of the focus error type** across all tested MDQFES models on the DUC dataset. This result also holds for the proprietary ESS dataset.

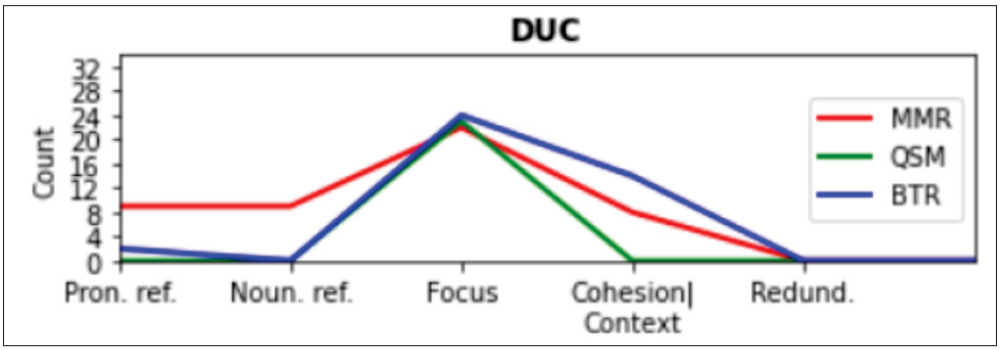


Figure 4.1 Qualitative study of prominent error types in system summaries MMR stands for Maximal Marginal Relevance, QSM for QuerySum, and BTR for Biased TextRank.

4.3 Sentence Encoders

Table 4.1 presents ROUGE scores for the BTR model on the ESS dataset across various sentence encoders from *HuggingFace*[4]. The first subtable reports scores for the encoder used in the original BTR implementation, the 2nd subtable for SBERT-based sentence encoders, and the last for asymmetric semantic search encoders. Bold font indicates the top score for each ROUGE variant.

Table 4.1 ROUGE scores for BTR on ESS across sentence encoders

Sentence encoder	R-1	R-2	R-L	R-SU4
bert-base-nli-mean-tokens	22.91	0.07	19.06	10.02
xlm-r-distilroberta-base-paraphrase-v1	**34.95**	**19.28**	**31.42**	**18.78**
paraphrase-xlm-r-multilingual-v1	34.90	19.24	31.37	18.74
distilroberta-base-paraphrase-v1	33.97	18.30	30.24	17.93
paraphrase-distilroberta-base-v1	34.00	18.34	30.27	17.95
msmarco-distilroberta-base-v2	33.52	18.16	30.28	17.44
msmarco-roberta-base-v3	34.56	19.10	31.25	18.44

[4] https://huggingface.co/

4.4 Query Formulation

See sections 6.5.3 and table 6.1, respectively, for the experiment report of Multi-Bias TextRank (MBTR) with query expansion and its results.

Table 4.2 presents ROUGE scores for the varied reductions experiment on the MBTR model. Section 6.6 and table 6.1 motivate the α and β choices.

Table 4.2 ROUGE scores for various reductions on Multi-Bias TextRank (MBTR). Bold font denotes the top scores in each subtable

Experiment	Reduction strategy	R-1	R-2	R-L	R-SU4
ERT[a] $\rightarrow$ MBTR$\vert_{\alpha=0.1,\beta=0.1}$	**Summation**	**45.51**	**28.22**	**41.61**	**28.11**
	Max	39.86	23.23	36.22	22.57
	Mean	40.95	24.09	37.17	23.81
	Median	40.24	23.70	36.35	23.54
	Conjoint probabilities	45.42	28.15	41.53	28.05
	Negative variance	23.16	9.75	19.96	10.48
SB[b] $\rightarrow$ MBTR$\vert_{\alpha=0.1,\beta=0.2}$	Summation	44.11	25.77	39.58	25.64
	Max	41.73	23.81	36.84	24.17
	Mean	41.87	24.04	36.97	24.27
	Median	42.07	24.12	37.18	24.38
	Conjoint probabilities	**44.13**	**25.80**	**39.61**	**25.67**
	Negative variance	31.63	16.33	27.64	16.58

[a] ERT is an input compound bias and stands for Expanded Reference-Terms (6.5.3)
[b] SB is an input compound bias and stands for Sentiment Biases (6.5.4)

4.5 Solutions at the ESS level

Sections 6.5.3, 6.5.4, and table 6.1 respectively present: a use of Multi-Bias TextRank (MBTR) with ESS-specific query expansion; a use of MBTR with sentiment biases and sentiment-based query expansion; ROUGE scores of the conducted experiments over MBTR and baseline models.

CHAPTER 5

DISCUSSION

Chapter 4 presented the experimental results, which we discuss and interpret in this chapter.

5.1 Qualitative study of prominent error types in system summaries

The finding of "focus" as the most impactful error type in Query-Focused Summarization is intuitive; a non-query-focused summary will have an almost null overlap with its reference and thus produce extremely low ROUGE scores. Other error types (e.g. redundancy, broken anaphoric references, cohesion) partially impact the summary's quality, which contrasts with the comprehensive influence of focus errors.

5.2 Sentence Encoders

The hypothesis of asymmetric semantic search encoders being more appropriate for our QFS task failed. This is possibly due to our problem space of short summaries, given that our ESS dataset presents single-sentence reference summaries (see sections 4.1 and 4.2 in chapter 6). In the conclusion section of chapter 6, we suggest re-testing such encoders in use cases with longer reference summaries.

The *xlm-r-distilroberta-base-paraphrase-v1* encoder performed best across all ROUGE scores. We surmise this to be the case due to its powerful combination of a distilled RoBERTa (Liu *et al.*, 2019) language model and its cross-lingual pre-training on the paraphrasing task enabling a deep understanding of language patterns.

5.3 Query Formulation

See section 6.6 for results and discussion of the query formulation experiments.

Regarding the reduction experiments, we re-iterate ROUGE-SU4 (6, section 4.2) as our variant of choice for performance comparison of the results in table 4.2. It is noteworthy as anecdotal

evidence that all ROUGE variants followed ROUGE-SU4 regarding best scores across all of our experiments.

For the ERT experiment, **mean** and **median** reductions present a 1.14% relative difference in performance and 0.45% in the case of the SB experiment. These marginal differences between the mean and median indicate low outlier biases in our query combinations. This is an intuitive finding, given that queries in the ESS task are expected to have a common objective rather than divergent ones.

The **negative variance** reduction produced the worst results for both the ERT and the SB experiments. This disproves the hypothesis of bias accordance being a good criterion for query-relevance scoring.

Summation performed 21.86% better than the **max** reduction in the ERT experiment, and 5.9% in the SB experiment. This finding highlights the importance of combining all biased contributions rather than selecting one and nullifying the contributions of the rest.

Summation performed best in the ERT experiment. In the SB experiment, summation performed 0.12% worst than **conjoint probabilities**. We elect to trade off this marginal difference for the uniform choice of reduction strategy across query combinations, i.e. summation, and for the lower computational cost compared to the sum of logarithms implementation of conjoint probabilities.

5.4 Solutions at the ESS level

See section 6.6 for results and discussion of the experimental solutions at the ESS level.

CHAPTER 6

QUERY-FOCUSED EXTRACTIVE SUMMARIZATION FOR SENTIMENT EXPLANATION

Ahmed Moubtahij[a] , Sylvie Ratté[a] , Yazid Attabi[b] , Maxime Dumas[b]

[a] Department of Systems Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[b] Croesus, 600 Bd Armand-Frappier, Laval, Québec, Canada H7V 4B4

Paper submitted for publication, March 2023

6.1 Abstract

Constructive analysis of feedback from clients often requires determining the cause of their sentiment from a substantial amount of text documents. To assist and improve the productivity of such endeavors, we leverage the task of Query-Focused Summarization (QFS). Models of this task are often impeded by the linguistic dissonance between the query and the source documents. We propose and substantiate a multi-bias framework to help bridge this gap at a domain-agnostic, generic level; we then formulate specialized approaches for the problem of sentiment explanation through sentiment-based biases and query expansion. We achieve experimental results outperforming baseline models on a real-world proprietary sentiment-aware QFS dataset.

6.2 Introduction

Sentiment analysis is the Natural Language Processing (NLP) task of predicting the affective state of a text passage. It is generally useful for applications concerned with feedback analysis of experiences (e.g., products, events, or services). However, simply being aware of the sentiment does not enable improvement of the experience; this purpose requires knowledge of the specific causes and features related to the sentiment.

Given a multitude of documents, a sentiment of interest (e.g., negative or positive), and a query regarding the targeted entities (e.g., a specific product, date, or location), our main objective is to provide an informative summary of the input documents that explains the cause(s) of the queried sentiment. This goal falls under a constrained QFS task, which we term Explicative Sentiment Summarization (ESS). See Figure 6.2 for a depiction of this process.

Compared to the Question Answering task's factoid outputs, the QFS task is motivated by more complex and contextually rich responses. It is thus a more appropriate parent task for ESS, which consists of elaborating on the cause(s) of the queried sentiment. The problem space of ESS is marginally akin to that of the Aspect-Based Sentiment Analysis (ABSA) task. ABSA associates sentiments with specific aspects (categories, features, or topics). Such aspects are predefined or extracted by a pipeline component, and the sentiment of each is a prediction objective. ESS concerns use cases where the target sentiment is prior knowledge and is thus an input item. Leveraging the latter allows simplifications such as computing the strength of the targeted sentiment for each text passage, thus inherently circumventing aspect identification. Additionally, ABSA produces sentiment associations for each aspect, whereas ESS outputs a natural language summary explaining the cause of the queried sentiment.

A common shortcoming of the QFS task and its proposed models is the putative gap between the source text and the input query in terms of *Language Register* (LR, formality level) and *Information Content* (IC, from Shannon's Information Theory). An LR gap occurs when, for example, a colloquial query formulation addresses source text written in formal style or in domain-specific terminology. An IC gap is typically incurred by the generic semantic coverage of short queries in relation to the specific semantics in detailed source text passages.

Our following contributions first address this issue at a generic level, then at a specialized level for our purpose of sentiment explanation:

1. We introduce the *Compound Bias-Focused Summarization (CBFS)* (6.4.1) framework to improve the chances of aligning the user's intent with arbitrary and possibly heterogeneous language registers in source documents by supporting multiple query formulations;

2. We concretize the CBFS framework with our *Multi-Bias TextRank* (MBTR) (6.4.2) model and its *Information Content Regularization* (6.4.3) which guides the QFS process towards the desired level of specificity;
3. We introduce the *Explicative Sentiment Summarization (ESS)* task, (6.4.4) which specializes the QFS task by leveraging prior knowledge in a sentiment explanation setting;
4. We substantiate the ESS task with sentiment-based bias computation (6.4.4.2) and query expansion (6.4.4.3).

6.3 Related Work

The following is an overview of the literature relevant to our task and contributions, spanning works in query-focused extractive summarization and query expansion.

6.3.1 Query-Focused Extractive Summarization

The NLP task of automatic summarization aims to compress a document or collection of documents into a salient and concise summary. Jones (1998) introduces three context factors concerned with automatic summarization and its evaluation: the nature of the input text (e.g., its domain and structure); the nature of the output summary; the purpose of the summary. Ter Hoeve *et al.* (2022), who ground their work in that of Jones (1998), advocate for the usefulness of a summary concerning the user's needs. They report that the purpose factors receive the least attention from works in automatic summarization, barring the latter's specializations which consider the audience and the situation. Among the latter is the task of Query-Focused Summarization, of which the expected output is a summary of the input document(s) that focuses on the query.

Automatic summarization can be achieved either by a semantic abstraction of the source text's salient information, or by a verbatim extraction of it.

While human-level summarization is abstractive, in practice, recent works (Ladhak, Durmus & Hashimoto (2022a); Ladhak, Durmus, He, Cardie & McKeown (2022b); Balachandran,

Hajishirzi, Cohen & Tsvetkov (2022); Fischer, Remus & Biemann (2022)) are still attempting to solve text generation errors such as factuality and hallucination. These shortcomings make abstractive summarization models currently unreliable in applications with tangible stakes.

Extractive summarization selects and concatenates salient text spans. This approach potentially hinders the cohesion of the summary as a whole. Indeed, text cohesion is a generally desired attribute and yet one of the most common error types in extractive summaries (Kaspersson *et al.* (2012); Smith *et al.* (2012)). However, it may be an optional attribute for critical applications prioritizing content reliability, output traceability, and fact-checking, all facilitated in extractive summarization.

Numeric representation of text is ancillary to the automatic summarization task, since it enables arithmetic transformations from the task's input space to its output space. Given the importance of pragmatics in natural language, the usefulness of such representations is greatly improved by their sensitivity to context. *BERT* Devlin *et al.* (2019), a pre-trained transformer Vaswani *et al.* (2017) encoder-based architecture, has seen widespread use as a Pre-trained Language Model (PLM) across recent text summarization systems (Liu & Lapata (2019); Laskar *et al.* (2020a); Kazemi *et al.* (2020); Laskar *et al.* (2020b); Xu & Lapata (2020); Xu & Lapata (2021); Xu & Lapata (2022); Laskar, Hoque & Huang (2022)). These models' State-Of-The-Art (SOTA) performance motivated us to adopt BERT-based models for text representation in automatic summarization.

Maximal Marginal Relevance (MMR) Carbonell & Goldstein (1998) is a Multi-Document Query-Focused Extractive Summarization (MDQFES algorithm that conjointly considers diversity and query relevance when retrieving salient passages from a collection of documents.

Liu & Lapata (2019) propose *BertSum*, a BERT-based model fine-tuned for both abstractive and extractive summarization, respectively, on the *XSUM* Narayan *et al.* (2018) dataset as *BertSumAbs*, and on the *CNN/DailyMail* Hermann *et al.* (2015) dataset as *BertSumExt*. Laskar *et al.* (2020a) pre-train BertSum similarly to BertSumAbs, then fine-tune it on the *DebatePedia* dataset Nema *et al.* (2017) for Query-Focused Extractive Summarization (QFAS).

Motivated by the success of BERT contextual embeddings, Kazemi *et al.* (2020)'s unsupervised *Biased TextRank* (BTR) model represents nodes from *TextRank* Mihalcea & Tarau (2004), a complete graph, as *SBERT* Reimers & Gurevych (2019) sentence encodings. BTR then subjects the underlying *PageRank* Page *et al.* (1999) centrality computation to a lower bound similarity, and to a query-bias for every sentence-node. Thereby ranking the input sentences by a conjunction of their centrality and query-bias.

Xu & Lapata (2020) argue that disjoining intra-document salience and query-relevance allows for separate modeling of the query and for summaries to address specific questions; this motivates their coarse-to-fine model, *QuerySum*, where text passages from the input documents are sequentially processed through query-relevant retrieval, followed by evidence estimation based on the Question-Answering (QA) task, and then by centrality-based re-ranking, i.e., salience for the surrounding text passages.

Laskar *et al.* (2020b), Xu & Lapata (2021), Xu & Lapata (2022) and Laskar *et al.* (2022) propose different approaches to address the prominent issue of lack of labeled QFS datasets.

Laskar *et al.* (2020b) opt for a distant and weakly supervised approach for generating weak (artificially generated) reference summaries from gold reference summaries through a pre-trained, RoBerta-based Liu *et al.* (2019) sentence-similarity model.

Assuming that generic (non-query-focused) summaries contain information on latent queries, Xu & Lapata (2021)'s *MARGE* model uses selective masking to reverse-engineer proxy queries, then pairs them with input sentences scoring high on *ROUGE* Lin (2004) (see 6.5.2). Thus, enabling weak supervision for ranking query-relevant sentences that are subsequently fed to a length-controllable QFAS model with optional user-query.

The *LQSum* model (Xu & Lapata, 2022), unlike MARGE, does not assume the target queries' length and content, nor does it require a development set. It achieve this by discarding the sequential query modeling approach, and replacing it with a zero-shot-capable alignment

between the source tokens and discrete latent variables. The latter are expressed by a binomial distribution indicating the query relevance belief of a source token.

6.3.2 Query Expansion

The dissonance between query and object signals motivates the NLP task of Query Expansion (QE), which is ancillary to downstream tasks such as QA, Information Retrieval (IR), or QFS. QE generally employs techniques such as re-weighting query terms and/or augmenting them with semantically related terms (Riezler, Vasserman, Tsochantaridis, Mittal & Liu (2007); Ganu & P. (2018); Zheng *et al.* (2020)).

Riezler *et al.* (2007)'s query expansion methods leverage Statistical Machine Translation (SMT) for paraphrasing and mapping to answer terms. While such back translation methods might somewhat preserve semantics, they are liable to lose the domain property of language, which disqualifies it from our need to bridge the language register gap between user-query and domain-specific documents. This particular discrepancy is observed by Ganu & P. (2018) in the search feature of their accounting software, in which users employ colloquial language to query the formal and financial text in their knowledge base. They address this problem with strategies for synonym substitution and expansion to nearest neighbor-embeddings, based on vocabulary from their hand-curated proprietary dataset. Albeit a valid approach for aligning the domain of query language, crafting a problem-specific lexicon requires seldom available human resources and expertise.

Zheng *et al.* (2020) further the motivation of QE with the issue of noisy query expansion, for which they propose *BERT-QE*, a three-step QE model in which initially ranked documents are:

1. Re-ranked on query-relevance with a BERT model pre-trained on the *MS MARCO* Bajaj *et al.* (2018) QA dataset;
2. Chunked into passages for relevance scoring with the model fine-tuned on a target dataset;
3. Re-ranked based on passage document-relevance and query-relevance.

Zheng *et al.* (2020)'s QE approach is restricted to IR as a downstream task by considering retrieval objects as entire documents, which does not directly accommodate our target task of QFES since it retrieves sentences.

Akin to the QE task, the Term Set Expansion (TSE) task consists of expanding members of a semantic class from a small seed set of terms. Kushilevitz *et al.* (2020) propose two TSE methods based on BERT used directly as a Masked Language Model (MLM): In *MPB1* (MLM-Pattern-Based), seed-terms are masked in sentences in which they occur (indicative patterns), then an MLM predicts the masks in their contexts, at which point the correctly predicted masks have their next best predictions elected for query expansion; *MPB2* circumvents out-of-vocabulary masked terms in indicative patterns by collecting single- and multi-token terms from similar patterns.

Kushilevitz *et al.* (2020)'s methods leverage an MLM's vocabulary for expanding seed terms in the context of the input text, which does not require a handcrafted lexicon, and helps align the source documents' language register with that of the expanded seed-terms. In our work, we need only consider seed terms as query terms to utilize these TSE methods for query expansion.

6.4 Methodology

We establish a framework for combining multiple queries, concretize it with our MBTR model, then subject the latter to information content regularization. We introduce the ESS task for sentiment explanation and employ corresponding techniques with reference-based query formulation, sentiment bias, and query expansion.

6.4.1 Compound Bias-Focused Summarization

To the best of our knowledge, all current QFS models consider a single input query. This design burdens the query's formulation by concisely targeting all information of interest at various scopes of variance and depth. Presented with such a challenge, all query formats (Xu & Lapata, 2021) face the following difficulties: natural language articulation must encompass the full

intent; keywords circumvent the syntactic constraints of natural language at the cost of its expressive flexibility (e.g., contextual disambiguation); albeit concise, the typical brevity of a title might limit the specificity of attainable information; a composite of the latter formats allows for trade-off balancing but incurs a non-trivial choice of representation to accommodate its syntactic heterogeneity effectively.

To tackle the aforementioned challenge, we propose *Compound Bias-Focused Summarization (CBFS)* (Figure 6.1). In this framework, the effects of multiple biases are combined through a reduction strategy[5] and input to a QFS model. We use the term "bias" as a generalization over query formats and non-query biases (e.g., 6.4.4.2). Providing multiple bias channels alleviates the burden in query formulation by partitioning the compromises mentioned above, instead of imposing them on a single query. Intuitively, this is analogous to humans reformulating questions from multiple perspectives or through various language registers for a wider coverage of their audience. Audience consideration is a heading of the advocated summarization purpose factor (Jones (1998); Ter Hoeve *et al.* (2022)).

6.4.2 Multi-Bias TextRank

Given its simplicity and flexibility, we extend Kazemi *et al.* (2020)'s BTR model to *Multi-Bias TextRank (MBTR)* to demonstrate the proposed CBFS framework.

Let n sentence encodings, d the embedding dimension, $\mathbf{b} \in \mathbb{R}^d$, $\mathbf{S} \in \mathbb{R}^{n \times d}$, α a control parameter, θ the similarity threshold and $\widetilde{\mathbf{A}} \in \mathbb{R}^{n \times n}$ a lower-bounded normalization of the weighted adjacency matrix $\mathbf{A} = sim(\mathbf{S}, \mathbf{S})$ such that:

$$\widetilde{\mathbf{A}}_{ij} = \begin{cases} \frac{\mathbf{A}_{ij}}{\sum_{j=1}^{n} \mathbf{A}_{ij}}, & \text{if } \sum_{j=1}^{n} \mathbf{A}_{ij} \neq 0 \text{ and } \frac{\mathbf{A}_{ij}}{\sum_{j=1}^{n} \mathbf{A}_{ij}} \geq \theta \\ 0, & \text{otherwise} \end{cases} \tag{6.1}$$

[5] (weighted) summation, max, conjoint probabilities, median, inverse variance, etc.

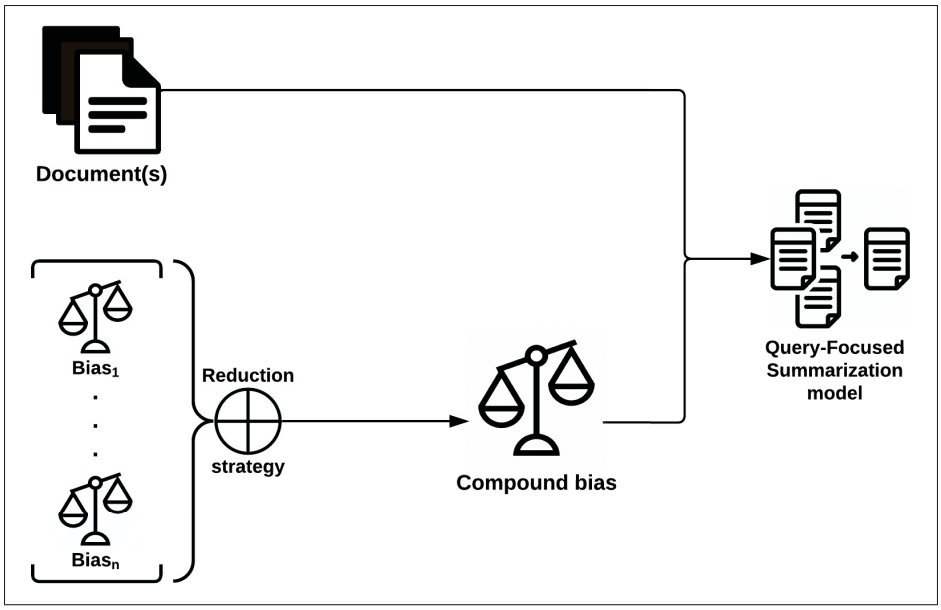


Figure 6.1 Compound Bias-Focused Summarization framework
The contributions of multiple biases are folded into a compound bias, which is then integrated into a Query-Focused Summarization model.

Then the PageRank vector in the Biased TextRank model can be recursively computed as follows:

$$R = \alpha\widetilde{\mathbf{A}}R + (1-\alpha)sim(\mathbf{b},\mathbf{S}) \tag{6.2}$$

Let q the number of query encodings, $\mathbf{B} \in \mathbb{R}^{q\times d}$ and μ a normalization function such as $\mu : \mathbb{R}^n \setminus \{\mathbf{u} : \mathbf{1}^\top\mathbf{u} = 0\} \to \mathbb{R}^n : \mathbf{u} \mapsto \mathbf{u}/(\mathbf{1}^\top\mathbf{u})$. Then the PageRank vector in our Multi-Bias TextRank model is expressed as follows:

$$R = \alpha\widetilde{\mathbf{A}}R + (1-\alpha)\mu\left(\bigoplus_{i=1}^{q} sim(\mathbf{B},\mathbf{S})_{i^*}\right) \tag{6.3}$$

We implement the $\oplus$ reduction operator as a summation, and the similarity function $sim : \mathbb{R}^{m\times k} \times \mathbb{R}^{n\times k} \to \mathbb{R}^{m\times n}$ as matrical cosine similarity:

$$sim(\mathbf{U}, \mathbf{V})_{ij} \coloneqq \frac{(\mathbf{U}\mathbf{V}^\top)_{ij}}{||U_{i*}|| \cdot ||V_{j*}||} \tag{6.4}$$

where $\coloneqq$ denotes "defined as", and the $i*$ and $j*$ subscripts denote a row-vector of a matrix. The μ normalization of the cumulative bias vector scales it comparably to the centrality vector $\widetilde{\mathbf{A}}R$.

While a single query formulation might not effectively address a desired sentence, folding the bias vectors of multiple queries increases its relevance score. Conjointly, a low score denotes more confidence in rejecting a sentence, given the implication that none of the query formulations neighbor it in semantic space.

The PageRank recursive term, R, in equations 6.2 and 6.3, computes centrality through the repeated transformation of itself by the weighted adjacency matrix $\widetilde{\mathbf{A}}$. Thus, R is essentially converging towards the eigenvector of $\widetilde{\mathbf{A}}$ with an eigenvalue of 1, i.e., the stationary probability distribution of the salience likelihoods of each sentence. Intuitively, this process simulates the broadcasting of sentence salience throughout the TextRank graph. In other words, it iteratively amplifies the scores of sentences similar to important sentences until convergence[6]. Once $\widetilde{\mathbf{A}}$'s equilibrium distribution is sufficiently stable, the sentences associated with the top probabilities are selected as the output summary.

6.4.3 Information Content Regularization

Amigó, Ariza-Casabona, Fresno & Martí (2022) call attention to the formal properties of text embeddings, based on the notion of Information Content (IC) from Shannon's Information Theory. One such property is the correspondence of IC with the vector norm of a text unit's embedding. We leverage this feature to disfavor candidate sentences by their distance from the targeted level of specificity.

[6] A set number of iterations and/or an ϵ error tolerance.

Let $\mathbf{G} \in \mathbb{R}^{m \times d}$ a matrix of m sentence-encodings from a guiding example summary, and $\mathbf{\Delta}_{\text{IC}} \in \mathbb{R}^n$ the observed-to-target IC distances:

$$\mathbf{\Delta}_{\text{IC}i} \coloneqq \left| \, ||\mathbf{S}_{i*}|| - avg(\, (||G_{j*}||)_j \,) \, \right| \tag{6.5}$$

where $avg : \mathbb{R}^n \rightarrow \mathbb{R}$ denotes a statistical average, which we define as the arithmetic mean $avg(\mathbf{u}) \coloneqq \bar{\mathbf{u}}$. Then, with β as a control parameter[7], we penalize every bias vector $sim(\mathbf{B}, \mathbf{S})_{i^*}$ in Equation 6.3 by its distance from the target IC (Equation 6.5):

$$R = \alpha \widetilde{\mathbf{A}} R + (1 - \alpha) \mu \left(\bigoplus_{i=1}^{q} (sim(\mathbf{B}, \mathbf{S})_{i^*} - \beta \mathbf{\Delta}_{\text{IC}}) \right) \tag{6.6}$$

The sentences associated with $\mathbf{G}$ can be provided by application-specific prior knowledge (see 6.5.3), in which case the target IC, i.e., $avg(\, (||G_{j*}||)_j \,)$ is embedded in the system, or by a user's example text to guide the desired level of specificity.

6.4.4 Explicative Sentiment Summarization

For sentiment explanation, we can disregard open-domain queries and specialize the QFS task for biases and queries that align with this objective. Additionally, we can leverage the prior knowledge of queries in a sentiment explanation setting. We introduce the task of *Explicative Sentiment Summarization (ESS)*.

6.4.4.1 Reference-based Query Formulation

For any sentiment-aware QFS dataset, its summary references are expected to explain the queried sentiments. We leverage this expectation to dispense users of query formulation by automating

[7] Note that BTR $\equiv$ MBTR$|_{q=1,\beta=0}$

it in the ESS model, thus reducing the user query's burden to merely mentioning the specific entities of interest, such as product names or dates, which can then be appended to the automated query or considered a separate query as per 6.4.1.

A simple heuristic for automating query formulation in an ESS setting would be selecting the *Frequent Reference-Words (FRW)* or *Frequent Reference-Phrases (FRP)* from the development split of the ESS dataset. This approach has the advantage of embedding common answer signals directly into the QFS bias.

6.4.4.2 Sentiment Bias

Unlike the QFS task, ESS can make assumptions about the query, such as the user's prior knowledge regarding the sentiment of interest. This allows an ESS model to adapt its query-relevance computation consequently.

Sentiment classifiers are trained to predict the perceived polarity of a text passage. The use case of sentiment explanation assumes prior knowledge of the sentiment of interest; we can thus utilize the prediction probability of this sentiment for every input sentence to construct a *sentiment bias vector*. However, the latter is potentially insufficient for the ESS task since it does not encode information regarding the targeted entities (e.g., product name) and should thus be used in combination with complementary query-biases (6.4.1), as exemplified in Figure 6.2.

This ESS-specific approach demonstrates a novel bias method that contrasts with the conventional query-sentence similarity computation in QFS.

6.4.4.3 Sentiment-based Query Expansion

In addition to enabling a sentiment bias vector (6.4.4.2), the prior knowledge in ESS can also be utilized for sentiment-based query expansion.

We propose using a hyperparameter pair of small sentiment phrases to select from for expansion, for example, "excellent service" and "poor experience". The suggested brevity is motivated by its

correlation with low Information Content (6.4.3), i.e., less specificity, which should broaden the reach for expansion in semantic space. We use phrases as text units instead of words to leverage the collocational properties of PLMs and thus enhance representation in semantic space.

Figure 6.2 depicts the ESS system, which, given an input sentiment:

1. Selects the corresponding integrated sentiment phrase;
2. Decomposes the input document(s) into phrases (see 6.5.4);
3. Retrieves the top K document-phrases[8] with the most cosine-similar encodings to the sentiment phrases. These encodings are produced with an asymmetric semantic search encoder [9] given the brevity of the sentiment-phrase.

This QE method does not require an external lexicon or knowledge base and inherently circumvents the typical linguistic dissonance between the query and the source document(s).

6.5 Experiments

We present the used dataset and the evaluation metric, then apply our proposed methods in two main experiments: *MBTR with query expansion*, which requires a development set, and *MBTR with sentiment*, which does not.

6.5.1 Dataset

We use a proprietary ESS dataset of which only metadata is disclosable. This dataset spans 950 ESS units, each containing:

- The name of the targeted entity;
- The sentiment of interest;
- 1 to 576 documents with a mean of 17 and variance of 38, with each document spanning 2 to 771 sentences with a mean of 24 and variance of 36;

[8] We use K=30

[9] https://www.sbert.net/examples/applications/semantic-search/README.html

- A single-sentence abstractive reference summary explaining the sentiment.

We conduct experiments using 75% of examples as a development set, and 25% as a test set.

6.5.2 Evaluation Metric for Automatic Summarization

We use the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) metric as it is the de-facto standard in automatic summarization. ROUGE varies strategies to quantify the n-gram overlap of the output text with its reference(s). Our ESS dataset presents single-sentence summaries of multiple documents; Lin (2004) report the ROUGE-{1, L, SU4, SU9} variants as most correlating with human judgment in the *problem space of short summaries*. However, Owczarzak *et al.* (2012) advocate for ROUGE-2-R, Rankel *et al.* (2013) for ROUGE-{3, 4}, and Graham (2015) for ROUGE-2-P.

Given the above discordance, we heuristically elect ROUGE-SU4 by the criterion of top variance through numerous experimental runs on our dataset, hypothesizing that high variance denotes reactivity to summary quality and low variance insensitivity to it; thus, we report *ROUGE-SU4*. We find its F1 score is also reported in recent works (Xu & Lapata (2020); Xu & Lapata (2021); Xu & Lapata (2022); Laskar *et al.* (2022)) in combination with *ROUGE-1, ROUGE-2* and *ROUGE-L* F1-scores (Laskar *et al.* (2020a); Kazemi *et al.* (2020)), which we also report using the *pythonrouge*[10] implementation.

6.5.3 Multi-Bias TextRank with Query Expansion

We use the NLTK[11] library to decompose the input documents into sentences, and an SBERT[12] encoder to represent them and the following expanded queries in MBTR$|_{\alpha=\beta=0.1}$ (Equation 6.6):

[10] https://github.com/tagucci/pythonrouge

[11] https://www.nltk.org/

[12] https://huggingface.co/sentence-transformers/xlm-r-distilroberta-base-paraphrase-v1

1. **FRW-MPB2**: we construct an FRW query with the top 20 frequent non-stopwords from the development set, then expand it with MPB2 (6.3.2), using its authors' (Kushilevitz *et al.*, 2020) reported hyperparameters;
2. **FRP-MPB2**: we redefine text units in FRW-MPB2 as noun phrases, which we obtain using the spaCy library's noun chunks feature[13];
3. **FRP-BTR**: we expand the FRP query using BTR (Kazemi *et al.*, 2020) with phrases as text units[14], then re-rank its output by descending frequency in the input documents and retrieve the top 20 phrases.

Before concatenating the individual terms (words or phrases) for each of the above query expansions, we remove duplicates, terms entirely composed of stopwords, and mentions of specific entities such as dates or organization names – using spaCy's NER[15] feature – to avoid spurious skewing towards a subset of the input sentences. We preserve Kazemi *et al.* (2020)'s recommended $\theta = 0.65$ for the similarity threshold (Equation 6.1) in all (M)BTR experiments.

The FRW-MPB2 + FRP-MPB2 + FRP-BTR query combination will hereafter be referred to as *Expanded Reference-Terms (ERT)*.

In the ESS task, we prepend the targeted entity's name to each query before and after expansion. Doing so produces deliberate skewing towards entity-relevant sentences. Additionally, we construct the $\mathbf{G}$ encodings matrix in Equation 6.5 from reference sentences in the development set.

6.5.4 Sentiment-aware Multi-Bias TextRank

Given input documents, a queried entity and sentiment, Figure 6.2 depicts the following process:

1. We use a sentiment classifier to predict the probability of the given sentiment for every input sentence, thus producing a *sentiment bias vector*;

[13] https://spacy.io/usage/linguistic-features#noun-chunks

[14] https://github.com/DerwenAI/pytextrank

[15] https://spacy.io/usage/linguistic-features#named-entities

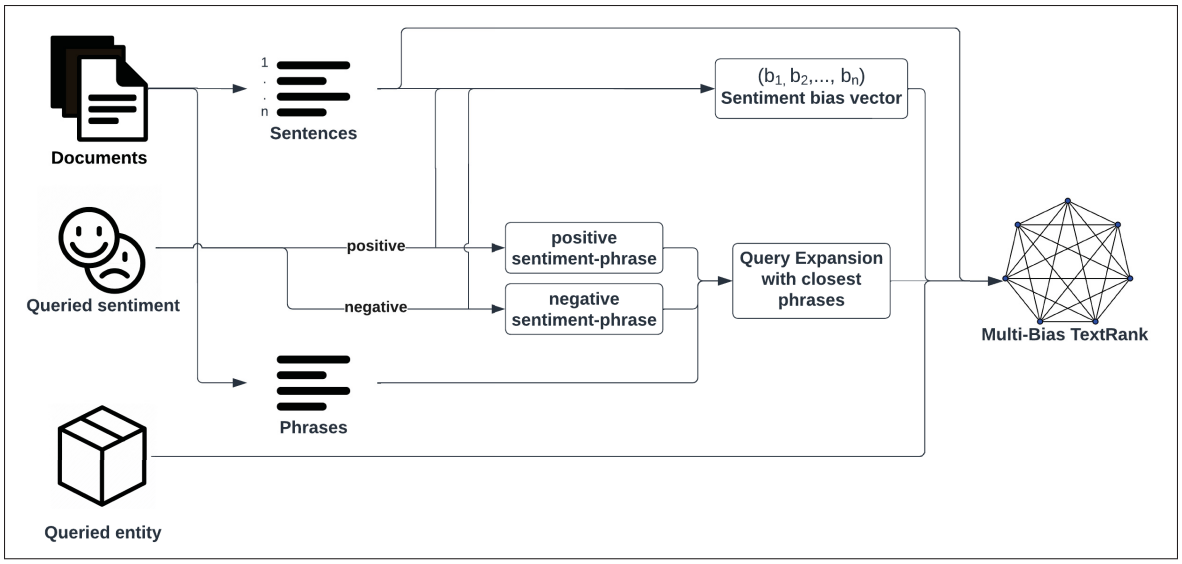


Figure 6.2 Explicative Sentiment Summarization system: integration of query expansion and sentiment bias into Multi-Bias TextRank.

2. We select the sentiment-corresponding query from a hyperparameter pair of sentiment phrases, then expand it to its top K[16] most cosine-similar document phrases in the space of an asymmetric semantic search encoder[17]. The resulting expanded queries are prepended with the queried entity;
3. We combine the sentiment bias vector with the expanded queries' bias vectors in $\text{MBTR}|_{\alpha=0.1,\beta=0.2}$ (Equation 6.6).

In the second step above, phrases are noun phrases (NP) and verb phrases (VP). NPs are extracted with spaCy's noun chunking feature, as declared in 6.5.3. We specialize VP patterns for the ESS task using spaCy's rule-based matching[18] such as:

```
0. vp_pattern = [
1. {},
2. {'POS': 'AUX', 'OP': '?'},
```

[16] We use K=30

[17] msmarco-distilbert-base-v4

[18] https://spacy.io/usage/rule-based-matching

```
3. {'DEP': 'neg', 'OP': '?'},
4. {'POS': 'VERB', 'OP': '+'},
5. {'POS': 'ADV', 'OP': '*'},
6. {'POS': 'ADJ', 'OP': '+'},
7.]
```

The numbered lines respectively describe: 1) a wildcard representing any token; 2) an optional auxiliary such as "is", "was", "could", or "should"; 3) an optional negation such as "not"; 4) at least one verb such as "trend", "trending", or "react"; 5) none or multiple adverbs such as "significantly"; 6) at least one adjective such as "worse" or "better". Thus, an example VP matching these rules could present as "*[entity] is trending significantly worse*".

The combination of the sentiment bias vector and the sentiment-based query expansion will hereafter be referred to as *Sentiment Biases (SB)*.

6.6 Results and discussion

Table 6.1 presents ROUGE scores of experiments partitioned across the following list of subtables:

1. The upper bound expresses the maximum achievable scores given that the references are abstractive summaries;
2. MMR, QuerySum, and BTR are used as baseline MDQFES models for comparison. $\text{BTR}|_{\alpha=0.1}$ performs best among baselines across all reported ROUGE variants;
3. Each query expansion from ERT (6.5.3) is tested individually on $\text{BTR}|_{\alpha=\{0.1,0.85\}}$. The FRW-MPB2 query performs best across all reported ROUGE variants;
4. $\text{MBTR}|_{\alpha=\{0,0.1\}\times\beta=\{0,0.1\}}$ is tested with ERT as input. $\text{MBTR}|_{\alpha=0.1,\beta=0.1}$ performs best across all reported ROUGE variants. It also outperforms BTR with each ERT query (subtable 3), thus demonstrating the benefit of CBFS; this holds even with ablation of the ICR component (6.4.3) with $\text{MBTR}|_{\alpha=0.1,\beta=0}$;

Table 6.1 ROUGE scores of our 6.5.3 and 6.5.4 experiments.
Bold font denotes each subtable's top ROUGE variant score.

α	β	Experiments	R-1	R-2	R-L	R-SU4
-	-	Upper bound[a]	72.86	48.60	72.05	49.63
-	-	BQ[b]→[c]MMR[d]	25.13	8.59	-	10.29
-	-	BQ→QuerySum	27.03	12.03	-	12.86
0.85	-	BQ→BTR[e]	31.98	16.91	28.61	16.69
0.1	-		**34.15**	**17.50**	**30.35**	**17.41**
0.85	-	FRW-MPB2→BTR	32.73	17.48	29.06	17.37
0.1	-		**41.67**	**24.50**	**37.69**	**24.35**
0.85	-	FRP-BTR→BTR	33.20	17.70	29.48	17.48
0.1	-		37.79	21.18	34.21	20.77
0.85	-	FRP-MPB2→BTR	31.97	17.12	28.50	16.78
0.1	-		38.57	22.32	34.98	21.76
0.1	0.1	ERT→MBTR	**45.51**	**28.22**	**41.61**	**28.11**
0	0.1		44.21	27.02	40.03	26.96
0.1	0		44.82	27.84	41.01	27.67
0.1	0.1	SB→MBTR	43.58	25.45	39.10	25.36
0.1	0		42.51	24.89	38.44	25.01
0.1	0.2		**44.11**	**25.77**	**39.58**	**25.64**
0	0.2		43.42	25.18	38.93	25.02

[a] Upper bound is computed by selecting the source sentence with the highest ROUGE-SU4 score (6.5.2).
[b] BQ denotes our baseline query: "Why did {queried entity}'s receive {positive, negative} feedback".
[c] We use the left-hand side of → to denote the query combination inputs.
[d] In MMR, sentence similarity is computed with spaCy's *en_core_web_lg* model.
[e] We use the same SBERT encoder (6.5.3) for BTR and MBTR

5. MBTR$|_{\alpha=\{0,0.1\}\times\beta=\{0,0.1,0.2\}}$ is tested with SB (6.5.4) as input. MBTR$|_{\alpha=0.1,\beta=0.2}$ performs best across all reported ROUGE variants.

Only the best-performing combinations of α and β are reported, in addition to combinations relevant to ablation studies.

Throughout all BTR and MBTR experiments, we observe that $\alpha = 0.1$ performs consistently better than Kazemi *et al.* (2020)'s recommended 0.85 and than the ablation of the centrality

component with $\alpha = 0$. This suggests that the solution space of ESS with short summaries (6.5.2) highly prioritizes query focus, without dropping the intra-document salience component since it helps elect the most central sentence among the most bias-relevant.

Dampening ICR performs best at $\beta = 0.1$ for the ERT experiments and at $\beta = 0.2$ for the SB experiments. Thus, for the problem space of ESS with short summaries, we recommend $\beta = 0.1$ when a development set is available for constructing the ERT queries, and $\beta = 0.2$ with SB otherwise. We interpret ERT's lesser regularization requirement as benefiting from its inherent proximity with the target specificity given its embedded answer signals (6.4.4.1).

CONCLUSION AND RECOMMENDATIONS

We approach the putative linguistic dissonance in the QFS task with the CBFS framework, which we concretize with the MBTR model. We then specifically address our purpose of sentiment explanation by introducing the ESS task and its system comprising sentiment-based biases and query expansions.

We find that the MBTR model significantly outperforms baseline QFES models and the BTR model it extends. In particular, given that we input the same queries individually to BTR, outperforming it substantiates the CBFS hypothesis of favoring desired sentences through multiple query formulations. Our results also indicate that the ESS task is more suitable than QFS when the query involves a known sentiment.

This work is limited by its focus on the problem space of single-sentence reference summaries and by its lack of testing on other ESS datasets. In future works, we plan on:

1. Adapting Aspect-Based Sentiment Analysis (ABSA) datasets to the ESS task. ABSA datasets consists of texts with sentiment-annotated aspects, thus, they are sentiment-aware and can be utilized to construct an ESS dataset. That is, an input sentiment and entity-query with the goal of identifying said aspects and their context as sentiment explanations;
2. Integrating other QFS models into the CBFS framework. Generally QFS models quantify the relevance of a text passage with respect to a single query. Whichever query-relevance mechanism they use, it can be compounded with other biases, as opposed to a single query's;
3. Asymmetric semantic search encoders, such as those we used for query expansion in ESS (6.5.4), might be better suited for the QFES process when the desired summaries are longer than one sentence;
4. Adapting re-ranking models such as BERT-QE (Zheng *et al.*, 2020) to sentences for integration into MBTR.

APPENDIX I

QUALITATIVE STUDY OF PROMINENT ERROR TYPES IN SYSTEM SUMMARIES

Table-A I-1 Summarization error types indices
For example, [0, 0, 4, 0, 3] signifies:
4 focus errors and 3 redundancy errors.

0	Broken pronominal anaphoric reference
1	Broken noun anaphoric reference
2	Focus
3	Lack of cohesion or context
4	Redundancy

1. DUC - MMR worst summaries sample

unit_id	query	extractive summary	RSU4-F	Error Distribution
d366i	What commercial applications exist for cyanide? What industries is it used in? How is it used? What are its potential dangers? What real damage has commercially-used cyanide caused?	**Newmont is about to begin a feasibility study to look at capital and operating costs to make sure that its process is commercially viable.** **In numerous instances, they have cited employers responsible for confined-space deaths by using the "general duty clause," a catchall provision of the federal Occupational Safety and Health Act designed to plug holes in instances where there has been a fatality or a serious life-threatening hazard.** **As a result, many illogical, discredited and potentially ruinous practices continue unabated.** **This helps create a positive image for the mining industry in its battles with environmentalists.** **'Security has been Newmont's prime concern, but it's just another risk in a risky business - and you resolve it by getting good people to protect you.'** **T. ferro-oxidans cannot cause disease because it can only develop on inorganic matter.** **And it is my duty and the obligation of Galactic Resources to make certain nothing detrimental happens to this historic treasure.** **Local groups have expressed fears of environmental damage from a possible escape of the cyanide solution used to leach the ore.** **Newmont officials say, however, that they are applying 'the same stringent precautions in Peru as we would in the state of Nevada'.** **His job, in the era between the end of World War II and 1976, was to simply cater to those who sought to use the desert for commercial gain.** **Like nearly everyone else in the business, he was convinced that the origins of what he had stolen were untraceable once identification marks had been removed.**	1.181	[4, 6, 11, 3, 0]
d301i	Identify and describe **types** of **organized crime** that **crosses borders** or involves more than one country. Name the countries involved. Also identify the **perpetrators involved** with each type of crime, including both individuals and organizations if possible.	**Though his speech was a model of diplomatic balance, he stressed the many mistakes made by US law enforcement agencies in the original response to the growth of organised crime which 'permitted the expansion of a powerful, well-organised crime syndicate'.** **The FBI chief, on a 10-day tour of eastern Europe, told a packed hall at the interior ministry police academy that the fact that Russian crime groups could 'use their existing and expanding criminal networks to exploit weapons grade radioactive materials is sufficient to warrant international concern and action'.** **Despite attempts to disguise the real ownership, however, the accounts frequently contain a common element, such as an address or a telephone or facsimile machine number linked to cartel members, the Justice Department said in papers filed in U. S. District Court in Atlanta.** **Western drug officials also say that widespread corruption in the police force, customs service and even in the judiciary and the absence of tough conspiracy laws has protected Nigerians from being arrested and brought to trial.** **Officials defined organised crime as a group activity whose prime purpose is the achievement of large-scale profits through criminal acts, which is long-term and continuing and includes some form of self-perpetuating criminal structure.** **He** made extensive use for the first time of pentiti (informers who have plea-bargained, equivalent to a 'super-grass') and he was able to demonstrate that the Mafia was not a casual group of criminals but a complex organisation with a hierarchy of command linking the various families and their businesses.	1.53	[2, 2, 5, 3, 0]

2. DUC - QuerySum worst summaries sample

unit_id	query	extractive summary	RSU4-F	Error Distribution
d354c	What types of **dangers** do **journalists find themselves in** and how are these situations related to their work? Are journalists specifically targetted? Are they endangered by the type of situations they are covering or by the content of their reports?	serbia's president slobodan milosevic came under stinging attack yesterday over the arrest and beating of more than 20 opposition politicians and journalists in belgrade. **leaders of serbia's democratic opposition denounced the violent crackdown - in the wake of the ousting of yugoslav federal president dobrica cosic earlier this week - as 'a brutal crime'.** **mr cosic, too, breaking his silence, added to the criticism of the serbian leader, likening him to hitler.** **democratic politicians demanded the release of those arrested and the resignation of the interior minister.** **but their call for a mass demonstration outside the federal parliament yesterday went unheeded.** **the violence with which police broke up tuesday night's protest against the removal of mr cosic apparently succeeded in intimidating opposition supporters.** **serbia's democratic opposition has become increasingly marginal as radical nationalist leaders, backed by paramilitary groups, have gained ground.** **'there is a new ascendancy of the ex-communists. we are heading for worse troubles,' said mr borislav mihailovic, a member of depos, a coalition of democratic groups.** **on tuesday night police fired rubber bullets and tear-gas at several thousand people gathered outside parliament, the largest demonstration in belgrade for more than two years.** **one police officer died from gunshot wounds and more than 30 people were injured in street battles. mr vuk draskovic, head of the serbian renewal movement (spo), the largest opposition party, was in serious condition with a fractured jaw and other injuries after he was beaten by police who burst into the party offices early yesterday morning, an spo spokesman said. mr draskovic, a charismatic figure who led violent demonstrations against the milosevic government in 1991, was arrested with senior party officials and local journalists.** **mr nicola milosevic, head of the serb liberal party, said it was 'an unprecedently violent action instigated by slobodan milosevic and his police. it is a totalitarian regime that arrests the leader of the biggest opposition party'. commenting on his ousting by radical serbs backed by socialists in the federal parliament, mr cosic accused president milosevic of 'despotic self-will' and a 'classical stalinist act'. in an allusion to adolf hitler, mr cosic described the serbian president, a former ally, as 'fuhrer of the serbian radical party'. bosnian serb troops were reported to be closing in on the moslem enclave of gorazde yesterday, as an upsurge of fighting threatened to disrupt aid operations across the former yugoslav republic, reuter reports from sarajevo. the un yesterday halted relief flights to sarajevo after snipers wounded four french soldiers and a us relief aircraft was hit by a bullet. a belgian tv cameraman was killed by a sniper in central bosnia.**	1.465	[0, 0, 10, 0, 0]
d431h	What new applications of **robot technology** are **in** current **use successfully**?	squat and deceptively awkward-looking, the andros 5a -- a state-of-the-art robot -- rode into camarillo on monday to show county law enforcement authorities what it can do. in a demonstration before members of the ventura county sheriff's department bomb squad, the three-foot-tall, remote-controlled robot lurched through a range of exercises at the sheriff's shooting range at the camarillo airport. the andros is the first such robot to be demonstrated in ventura county, where sheriff's officials hope to purchase the model or one similar to it. bomb squads in los angeles, orange county and san diego already have bomb-handling robots, authorities said. in those areas, the robots have proved invaluable in handling suspicious-looking devices and have become standard safety equipment for bomb squads. **"just look at him," the voice sneers, finally. "he doesn't like you at all." ouch. nothing in the rock circus prepares visitors for the sting of learning that rotten never liked them. but at least that is the only painful moment during a tour of the combination wax museum-robotics show-rock hall of fame. that is, unless you consider it painful to watch a bob dylan robot in ray-bans strumming its way through "the times they are a-changin'." created by the tussaud's group, the same company behind mme. tussaud's waxworks, rock circus features dozens of sculpted rock stars. each figure is accompanied by a snatch of music and a dash of narrated comment.**	2.511	[0, 0, 3, 0, 0]
d374a		**mr gary becker, named yesterday as this year's winner of the nobel prize for economics, is proof that economists have more to offer than dubious forecasts, indecipherable equations and contradictory conclusions about the behaviour of money and markets. revered among his fellow professionals for his seminal work on the economics of discrimination and human capital theory, the 61-year-old professor at the university of chicago has spent the past 40 years extending the discipline of economics far beyond the world of international trade and finance. mr becker was the first to show that racial discrimination was economically costly and relied upon the existence of monopoly power. only companies with market power could afford to ignore qualified candidates or refuse paying customers because of their colour or religion. so encouraging competition is the best way to end discrimination. he also investigated how we punish criminals and why prison sentences for unarmed robbery are much lower than those for armed robbery. mr becker argued that the criminal balanced the rewards from crime against the probability and cost of capture, so not all crimes should receive too tough a sentence. attaching a life sentence to unarmed robbery would reduce the number of thefts, but it might also increase the number of murders once the marginal disincentive for petty thieves to use guns was removed. mr becker has also examined the loosening of us divorce laws. contrary to conventional wisdom, allowing one party to file a divorce suit rather than requiring mutual consent has not led to more divorces in states which have taken this route. however, as mr becker predicted, the size of financial settlements has fallen sharply now that there is no need for an agreement to the divorce. mr lawrence summers, an economics professor at harvard university and currently chief economist at the world bank, says that policymakers should have predicted this result. 'one of the most important unintended consequences of the divorce revolution in the us - the impoverishment of hundreds and thousands of children - would have been apparent if they had studied becker's seminal approach.'**	2.656	[0, 0, 10, 0, 0]

3. DUC - Biased TextRank worst summaries sample

unit_id	query	summary	RSU4-F	Error distribution
d354c	What types of **dangers** do **journalists find themselves in** and how are these situations related to their work? Are journalists specifically targetted? Are they endangered by the type of situations they are covering or by the content of their reports?	**Reacting to mounting concern about atrocities, it warned that individuals violating human rights would be held responsible for their actions.** **But the diplomats cautioned that the actual use of force would complicate the conflict.** **And in each case there were visible strains on its cohesion, as leading member governments argued among themselves, and with the secretariat, over the course of action to be pursued.** **The violence is prompting the international community to look at Kashmir with renewed urgency, out of concern for the fighting and for the damage done to relations between India and neighbouring Pakistan.** **"People are scared," said Carlos Rodolfo Alvarado, an election official.** **Increasingly, they insist on the hard questions, particularly ones that challenge an authority's conduct if it appears to have no democratic base and abuses the rights of those with less influence.** **Not least, it has affected journalists who occupy that highly sensitive and fine-tuned midway point between politics and literature and have always wanted to believe that the pen is mightier than the sword.** The morbidity figures seem to suggest that journalists have pushed the frontiers of free reporting so far that they have made themselves frighteningly vulnerable. **Anybody who criticizes the people at the top is running a big risk."** "It appears to us that **these actions** are an attempt to intimidate members of the foreign press and prevent us from carrying out legitimate reporting activities in China," the correspondents' letter of protest said. **Castaneda's articles have usually been critical of the Salinas government.**	2.914	[1, 0, 7, 2, 0]
d442g	What are some outstanding instances of **heroic acts** when the **hero was in danger of losing his/her life while saving others in imminent danger** of losing their own lives? Instances should include events occurring in the **U. S. only**.	**The events were a shock to us all, but you acted quickly and without fear for your own safety.** **"They were lucky the dumpster sunk away from the power line" or they would have been shocked, too, she added.** **An electrical or gas line malfunction most likely started the fire, authorities said.** **Insurance for Courage He also said he wanted to make sure heroes or their survivors would not "suffer pecuniarily" as a result of their deeds.** **A rescue of a family member also is excluded, unless the rescuer was severely injured or killed.** **Above all, the saved person must have been in imminent danger of losing his life, and the hero must have risked his life in performing the rescue.** **Surprisingly, perhaps, a nominee can still be chosen for the honor even if the life-saving attempt failed.** **We give awards to those who risk their own lives to save others."** **Acting by Instinct To Carnegie hero Robert Jameson, a heroic act is done by instinct.** **The awards are given to people who risk their lives attempting to save others from fiery deaths, said Judith Copeland, executive director of the Burn Institute** **The Medal of Valor is given for an extraordinary act of heroism by a state employee, undertaken at great risk in an effort to save human life.** **"His death is not fair because he was so good, and he was only trying to help.** **"Bibbie intervening at that particular time perhaps saved Bernard's life," the sheriff's spokesman said.** **The awards honor people who have risked -- or lost -- their lives to save others.**	3.306	[1, 0, 9, 5, 0]

LIST OF REFERENCES

Amigó, E., Ariza-Casabona, A., Fresno, V. & Martí, M. A. (2022). Information Theory–based Compositional Distributional Semantics. *Computational Linguistics*, 48(4), 907–948. doi: 10.1162/coli_a_00454.

Badrinath, R., Venkatasubramaniyan, S. & Veni Madhavan, C. E. (2011). Improving Query Focused Summarization Using Look-Ahead Strategy. *Proceedings of the 33rd European Conference on Advances in Information Retrieval - Volume 6611*, (ECIR 2011), 641–652. doi: 10.1007/978-3-642-20161-5_64.

Bahdanau, D., Cho, K. & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv:1409.0473 [cs, stat]*. Retrieved from: http://arxiv.org/abs/1409.0473.

Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S. & Wang, T. [arXiv:1611.09268 [cs] version: 3]. (2018). MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. arXiv. Retrieved on 2023-02-08 from: http://arxiv.org/abs/1611.09268.

Balachandran, V., Hajishirzi, H., Cohen, W. W. & Tsvetkov, Y. (2022). Correcting Diverse Factual Errors in Abstractive Summarization via Post-Editing and Language Model Infilling. Retrieved on 2023-02-02 from: https://arxiv.org/abs/2210.12378v2.

Banerjee, S. & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72. Retrieved from: https://aclanthology.org/W05-0909.

Cao, M., Dong, Y., Wu, J. & Chi Kit Cheung, J. (2020). Factual Error Correction for Abstractive Summarization Models. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Retrieved from: https://www.aclweb.org/anthology/2020.emnlp-main.506.

Carbonell, J. & Goldstein, J. (1998). The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (SIGIR '98), 335–336. doi: 10.1145/290941.291025.

Demszky, D., Guu, K. & Liang, P. [arXiv:1809.02922 [cs]]. (2018). Transforming Question Answering Datasets Into Natural Language Inference Datasets. arXiv. Retrieved on 2023-01-20 from: http://arxiv.org/abs/1809.02922.

Denkowski, M. & Lavie, A. (2014). Meteor Universal: Language Specific Translation Evaluation for Any Target Language. *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 376–380. doi: 10.3115/v1/W14-3348.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

Durmus, E., He, H. & Diab, M. (2020). FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5055–5070. doi: 10.18653/v1/2020.acl-main.454.

Erkan, G. & Radev, D. R. (2004). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22, 457–479. doi: 10.1613/jair.1523. ZSCC: 0003019.

Fischer, T., Remus, S. & Biemann, C. (2022). Measuring Faithfulness of Abstractive Summaries. *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pp. 63–73. Retrieved from: https://aclanthology.org/2022.konvens-1.8.

Ganu, H. & P., V. D. (2018). Fast Query Expansion on an Accounting Corpus using Sub-Word Embeddings. *Proceedings of the Second Workshop on Subword/Character LEvel Models*, pp. 61–65. doi: 10.18653/v1/W18-1208.

Gao, Y., Zhao, W. & Eger, S. (2020). SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1347–1354. doi: 10.18653/v1/2020.acl-main.124.

Graham, Y. (2015). Re-evaluating Automatic Summarization with BLEU and 192 Shades of ROUGE. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 128–137. Retrieved from: http://aclweb.org/anthology/D15-1013.

Guo, Y. & Hu, J. (2019). Meteor++ 2.0: Adopt Syntactic Level Paraphrase Knowledge into Machine Translation Evaluation. *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 501–506. doi: 10.18653/v1/W19-5357.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. doi: 10.1109/CVPR.2016.90.

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M. & Blunsom, P. (2015). Teaching Machines to Read and Comprehend. *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, (NIPS'15), 1693–1701. Retrieved from: http://dl.acm.org/citation.cfm?id=2969239.2969428.

Hochreiter, S. & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. doi: 10.1162/neco.1997.9.8.1735. ZSCC: 0059290.

Jones, K. S. [arXiv:cmp-lg/9805011 version: 1]. (1998). Automatic summarising: factors and directions. arXiv. Retrieved on 2023-02-01 from: http://arxiv.org/abs/cmp-lg/9805011.

Kaspersson, T., Smith, C., Danielsson, H. & Jönsson, A. (2012). This also affects the context - Errors in extraction based summaries. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 173–178. Retrieved from: http://www.lrec-conf.org/proceedings/lrec2012/pdf/776_Paper.pdf.

Kazemi, A., Pérez-Rosas, V. & Mihalcea, R. (2020). Biased TextRank: Unsupervised Graph-Based Content Extraction. *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 1642–1652. doi: 10.18653/v1/2020.coling-main.144.

Khalid, U., Beg, M. O. & Arshad, M. U. [arXiv:2102.11278 [cs] version: 1]. (2021). RUBERT: A Bilingual Roman Urdu BERT Using Cross Lingual Transfer Learning. arXiv. Retrieved on 2023-03-17 from: http://arxiv.org/abs/2102.11278.

Kryściński, W., McCann, B., Xiong, C. & Socher, R. (2020). Evaluating the Factual Consistency of Abstractive Text Summarization. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Retrieved from: https://www.aclweb.org/anthology/2020.emnlp-main.750.pdf.

Kushilevitz, G., Markovitch, S. & Goldberg, Y. (2020). A Two-Stage Masked LM Method for Term Set Expansion. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6829–6835. doi: 10.18653/v1/2020.acl-main.610.

Ladhak, F., Durmus, E. & Hashimoto, T. (2022a). Tracing and Removing Data Errors in Natural Language Generation Datasets. Retrieved on 2023-02-02 from: https://arxiv.org/abs/2212.10722v1.

Ladhak, F., Durmus, E., He, H., Cardie, C. & McKeown, K. (2022b). Faithful or Extractive? On Mitigating the Faithfulness-Abstractiveness Trade-off in Abstractive Summarization. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1410–1421. doi: 10.18653/v1/2022.acl-long.100.

Laskar, M. T. R., Hoque, E. & Huang, J. (2020a). Query Focused Abstractive Summarization via Incorporating Query Relevance and Transfer Learning with Transformer Models. *Proceedings of the Canadian AI*, (Lecture Notes in Computer Science), 342–348. doi: 10.1007/978-3-030-47358-7_35.

Laskar, M. T. R., Hoque, E. & Huang, J. X. (2020b). WSL-DS: Weakly Supervised Learning with Distant Supervision for Query Focused Multi-Document Abstractive Summarization. *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5647–5654. doi: 10.18653/v1/2020.coling-main.495.

Laskar, M. T. R., Hoque, E. & Huang, J. X. (2022). Domain Adaptation with Pre-trained Transformers for Query-Focused Abstractive Text Summarization. *Computational Linguistics*, 48(2), 279–320. doi: 10.1162/coli_a_00434. Place: Cambridge, MA Publisher: MIT Press.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv:1910.13461 [cs, stat]*. Retrieved from: http://arxiv.org/abs/1910.13461. arXiv: 1910.13461.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, pp. 74–81. Retrieved from: https://aclanthology.org/W04-1013.

Liu, Y. & Lapata, M. (2019). Text Summarization with Pretrained Encoders. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3721–3731. doi: 10.18653/v1/D19-1387.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. [arXiv:1907.11692 [cs] version: 1]. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv. Retrieved on 2023-02-03 from: http://arxiv.org/abs/1907.11692.

Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4), 309–317. doi: 10.1147/rd.14.0309.

Mihalcea, R. & Tarau, P. (2004). TextRank: Bringing Order into Text. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404–411. Retrieved from: https://www.aclweb.org/anthology/W04-3252.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. J. (1990). Introduction to WordNet: An On-line Lexical Database*. *International Journal of Lexicography*, 3(4), 235–244. doi: 10.1093/ijl/3.4.235.

Narayan, S., Cohen, S. B. & Lapata, M. (2018). Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807. doi: 10.18653/v1/D18-1206.

Nema, P., Khapra, M. M., Laha, A. & Ravindran, B. (2017). Diversity driven attention model for query-based abstractive summarization. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1063–1072. doi: 10.18653/v1/P17-1098.

Nenkova, A. & McKeown, K. (2011). *Automatic Summarization*. doi: 10.1561/1500000015.

Otterbacher, J., Erkan, G. & Radev, D. R. (2009). Biased LexRank: Passage retrieval using random walks with question-based priors. *Information Processing & Management*, 45(1), 42–54. doi: 10.1016/j.ipm.2008.06.004. ZSCC: 0000113.

Over, P. D., Dang, H. T. & Harman, D. K. (2007). DUC in Context. Retrieved from: https://www.nist.gov/publications/duc-context. ZSCC: 0000213 Last Modified: 2017-02-17T13:33-05:00.

Owczarzak, K., Conroy, J. M., Dang, H. T. & Nenkova, A. (2012). An Assessment of the Accuracy of Automatic Evaluation in Summarization. *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pp. 1–9.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. [Techreport]. Retrieved on 2018-11-13 from: http://ilpubs.stanford.edu:8090/422/.

Palmer, M., Gildea, D. & Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1), 71–106. doi: 10.1162/0891201053630264.

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2001). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, pp. 311. doi: 10.3115/1073083.1073135.

Quanye, J., Rui, L. & Jianying, L. (2021). Using Query Expansion in Manifold Ranking for Query-Oriented Multi-Document Summarization. *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pp. 940–951. Retrieved from: https://aclanthology.org/2021.ccl-1.84.

Rajpurkar, P., Zhang, J., Lopyrev, K. & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392. doi: 10.18653/v1/D16-1264.

Rajpurkar, P., Jia, R. & Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789. doi: 10.18653/v1/P18-2124.

Rankel, P. A., Conroy, J. M., Dang, H. T. & Nenkova, A. (2013). A Decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 131–136. Retrieved from: https://aclanthology.org/P13-2024.

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992. doi: 10.18653/v1/D19-1410.

Riezler, S., Vasserman, A., Tsochantaridis, I., Mittal, V. & Liu, Y. (2007). Statistical Machine Translation for Query Expansion in Answer Retrieval. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 464–471. Retrieved from: https://aclanthology.org/P07-1059.

Schuster, M. & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673–2681. doi: 10.1109/78.650093. ZSCC: 0006172.

Sellam, T., Das, D. & Parikh, A. (2020). BLEURT: Learning Robust Metrics for Text Generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892. doi: 10.18653/v1/2020.acl-main.704.

Smith, C., Danielsson, H. & Jönsson, A. (2012). Cohesion in Automatically Created Summaries. *Proceedings of the Fourth Swedish Language Technology Conference*. Retrieved from: https://www.ida.liu.se/~arnjo82/papers/sltc-12-sdj.pdf.

Sparck Jones, K. (1972). A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL. *Journal of Documentation*, 28(1), 11–21. doi: 10.1108/eb026526.

Ter Hoeve, M., Kiseleva, J. & Rijke, M. (2022). What Makes a Good and Useful Summary? Incorporating Users in Automatic Summarization Research. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 46–75. doi: 10.18653/v1/2022.naacl-main.4.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, & Polosukhin, I. (2017). Attention is All you Need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 5998–6008). Curran Associates, Inc. Retrieved from: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Xu, X., Dušek, O., Li, J., Rieser, V. & Ioannis Konstas. (2020). Fact-based Content Weighting for Evaluating Abstractive Summarisation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. doi: 10.18653/v1/2020.acl-main.455.

Xu, Y. & Lapata, M. (2020). Coarse-to-Fine Query Focused Multi-Document Summarization. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3632–3645. Retrieved from: https://www.aclweb.org/anthology/2020.emnlp-main.296.

Xu, Y. & Lapata, M. (2021). Generating Query Focused Summaries from Query-Free Resources. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6096–6109. doi: 10.18653/v1/2021.acl-long.475.

Xu, Y. & Lapata, M. (2022). Document Summarization with Latent Queries. *Transactions of the Association for Computational Linguistics*, 10, 623–638. doi: 10.1162/tacl_a_00480. Place: Cambridge, MA Publisher: MIT Press.

Ya, J., Liu, T., Cao, J. & Guo, L. (2021). Heterogeneous Graph Neural Networks for Query-focused Summarization. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)* (pp. 720–728). Society for Industrial and Applied Mathematics. doi: 10.1137/1.9781611976700.81.

Yeh, Y.-T. & Chen, Y.-N. (2019). QAInfomax: Learning Robust Question Answering System by Mutual Information Maximization. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3370–3375. doi: 10.18653/v1/D19-1333.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. *arXiv:1904.09675 [cs]*. Retrieved from: http://arxiv.org/abs/1904.09675. arXiv: 1904.09675.

Zheng, H. & Lapata, M. (2019). Sentence Centrality Revisited for Unsupervised Summarization. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6236–6247. doi: 10.18653/v1/P19-1628.

Zheng, Z., Hui, K., He, B., Han, X., Sun, L. & Yates, A. (2020). BERT-QE: Contextualized Query Expansion for Document Re-ranking. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4718–4728. doi: 10.18653/v1/2020.findings-emnlp.424.