HEC MONTRÉAL
École affiliée à l'Université de Montréal

Information bias for unsupervised document summarization

par
Florian Carichon

Thèse présentée en vue de l'obtention du grade de Ph. D. en administration
(spécialisation Science des données)

Décembre 2023

HEC MONTRÉAL
École affiliée à l'Université de Montréal

Cette thèse intitulée :

Information bias for unsupervised document summarization

Présentée par :

Florian Carichon

a été évaluée par un jury composé des personnes suivantes :

Laurent Charlin
HEC Montréal
Président-rapporteur

Gilles Caporossi
HEC Montréal
Directeur de recherche

Jian Tang
HEC Montréal
Membre du jury

Amine Trabelsi
Université de Sherbrooke
Examinateur externe

Firstname Lastname
HEC Montréal
Représentant du directeur de HEC Montréal

Résumé

Un résumé de document est une forme abrégée d'un contenu qui préserve son information principale afin d'être pertinent pour répondre à une tâche et à une audience particulière. Avec l'augmentation accélérée de la quantité de données sur internet, et notamment de texte, il devient essentiel de proposer des méthodes automatiques permettant de rendre digeste cette information. Dans ce contexte, les méthodes non supervisées, ne nécessitant aucun travail humain d'étiquetage, deviennent des outils salvateurs et puissants pour le résumé de document. Afin de créer un résumé, il est indispensable de considérer différents facteurs influençant la production d'un texte. On pense régulièrement à la donnée fournie aux modèles ou à la forme directe du résumé, mais il existe des facteurs plus implicites tels que les facteurs d'intention comme l'utilisation et l'auditoire auxquels le résumé est consacré. Au cours de nos travaux, nous nous sommes intéressés au rapport apparaissant entre ces facteurs contextuels et la perception et la caractérisation de la pertinence de l'information. Nous avons notamment fait le lien entre l'utilisation indicative ou informative d'un résumé avec la pertinence de sujet qui peut être sélective ou centrale. Nous avons aussi établi une connexion entre audience spécifique et générique avec la pertinence de nouveauté diverse ou redondante. Nous avons voulu mettre en avant la manière dont ces différentes intentions pouvaient influencer les méthodes pour représenter, marquer et produire de l'information. Pour cela, nous avons tout d'abord retracé l'historique des approches non supervisées existantes afin de proposer une typologie les étudiant sous cet angle de la pertinence d'information. Ensuite, nous avons appliqué cette vision au sein de trois articles répondants aux besoins de mise à jour, d'objectivité, ou de diversification de l'information.

Dans chacun de ces contextes, nous avons démontré que prendre en compte ces propriétés dans la caractérisation de la pertinence permettait d'améliorer les performances de nos algorithmes de référence. Nous avons également vérifié l'intérêt de bien qualifier ce besoin spécifique dans les jeux de données et les méthodes d'évaluation pour faire apparaître ces différentes notions fondamentales. Finalement, nous présentons une discussion afin d'analyser comment ce facteur impacte concrètement l'écosystème du résumé automatique de document. Ce travail permet de contribuer à la littérature en mettant en avant ce lien peu étudié avec théorie de l'information, et aussi de proposer de nouvelles pistes de recherches pour améliorer la compréhension des méthodes de traitement du langage naturel.

Mots-clés

Intelligence Artificielle; Traitement automatique du langage naturel; Résumé automatique de documents; Approches non supervisées; Pertinence de l'information; Facteurs d'intention;

Méthodes de recherche

Thèse par article; Revue de littérature; Proposition théorique.

Abstract

A document summary is an abridged form of content that preserves its main information to be relevant to a particular task and audience. With the accelerating growth of data quantity on the Internet, particularly text, it is becoming essential to offer automatic methods for making this information digestible. In this context, unsupervised methods, which require no human labeling work, are becoming powerful, life-saving tools for document summarization. To create a summary, it is essential to consider the various factors influencing text production. We regularly think of the data provided to models or the direct form of the summary, but there are more implicit factors such as intention factors like the usage and audience to which the summary is dedicated. During our work, we have been interested in the relationship between these contextual factors and the perception and characterization of information relevance. Specifically, we have linked the indicative or informative utilization of a summary with topical relevance, which can be selective or central. We also established a connection between specific and generic audiences with the notion of novelty relevance which can be diverse or redundant. We wanted to highlight how these different intentions could influence methods for representing, scoring, and generating information. To do this, we first retraced the history of existing unsupervised approaches, to propose a typology studying them from this angle of information relevance. Then, we applied this vision to three articles responding to the needs of updating, objectivity, or diversification of information. In each of these contexts, we demonstrated that taking these properties into account in the characterization of relevance improved the performance of our reference algorithms. We have further verified the importance of properly qualifying this specific need in datasets and evaluation methods to bring out these different fundamental notions. Finally, we present a discussion to analyze how this factor concretely impacts the automatic document summarization ecosystem. This work allows us to contribute to the literature by highlighting this little-studied link with information theory, and to propose new avenues of research to improve understanding of natural language processing methods.

Keywords

Artificial Intelligence; Natural language processing; Automatic document summarization; Unsupervised methods; Information relevance; Purpose factors.

Research methods

Thesis writtent with articles; Litterature review; Theoritical contribution.

Contents

List of Tables

List of Figures

List of acronyms

AI	Artificial Intelligence
NLP	Natural Language Processing
LLM	Large Language Model
LM	Language Model
CM	Constraint Model
MDS	Multi-Document Summarization
AE	Autoencoder
VAE	Variationnal Autoencoder
KL	Kullback-Leibler
MTL	Multi-task Learning
RNN	Recurrent Neural Network
GRU	Gated Recurrent Unit
GRL	Gradient Reversal Layer
TFIDF	Term Freauency Inverse Document Frequency
SVD	Singular Value Decomposition

LDA	Latent Dirichlet Allocation
MMR	Maximal Marginal Relevance
ROUGE	Recall Oriented Understudy for Gisting Evaluation
AMT	Amazon Mechanical Turk
DUC	Document Understading Conference
TAC	Text Analysis Conference
TREC	Text Retrieval Conference

Acknowledgements

This thesis would not have been possible without the support I have received and affection I have received along this journey. First and foremost, I would like to thank my family for putting up with and supporting me during this difficult and stressful period of my life. I am lucky to have my parents Martine and Bernard who have always allowed me to sharpen my curiosity and to be behind me to let me be serene and focused on this life goal. As for partner Yiu-Yin, I hope to show appreciation for you for your understanding, patience and unfailing love, which has kept me happy and myself during this period.

I want to warmly thank my thesis supervisor Gilles Caporossi for guiding me along the way. I was fortunate in that he could accompany me as I gained autonomy as a researcher, giving me the space to be independent and to make my own mistakes, which helped me to progress in life. At the same time, I would also like to express my gratitude Pr. Hongping Tan, Pr. Jian Tang, Pr. Bang Liu, and Pr. Denis Larocque for providing me with excellent feedback during the various stages of evaluating and writing this thesis.

I deeply thank to all my friends in Montreal and in France. I am thinking specifically of Johan Saba, Julien Silvio and Romain Rampa, Clément Jacquet and Nicolas Charlot who have been a breath of fresh air and a source of joy, laugh, and support for me. I would also wish to acknowledge all my HEC PhD colleagues and fellows, Patrick Groulx, Anaïs Galy, Francis Maisonneuve, Adrien Simonnot and Julie Gauneau, with whom I could collaborate, have fascinating discussions on several topics and spend some excellent evenings.

Finally, I would like to say a very special thank you to Florent Fettu and Chrys Ngouma, both M.Sc. students who participated in the programming of certain elements of

my articles and saved me precious time in completing my doctorate. More generally, I feel fortunate to have received technical support and funding from the Fédération des Caisses Desjardins du Québec and IVADO through the MITACS Acceleration grants IT16112. Special thanks to Étienne Plante-Dubé for giving me this opportunity and to Adel Benlagra for his technical support during the thesis. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

Preface

A thesis on natural language processing in 2023 that doesn't address the topic of large language models, what a heresy. Is it worth reading? Working on my thesis during this pivotal period in NLP has been interesting, rewarding but also so exhausting and emotionally stressful. Since the first publication of BERT in 2019 and GPT4 and ChatGPT in 2023, it was difficult as a new researcher to anticipate the meteoric emergence of this revolution for the field. This resulted in a new arms race to obtain evermore powerful systems where large teams and more modern machines were more and more needed. For a young researcher in a small laboratory, this left two viable options for research: find ways of applying these new models to improve performances on some tasks or identify ways of analyzing and comprehend these models. Knowing that I was already well advanced in my PhD, I had already developed a deep interest in unsupervised approaches and information perception in document summarization. I was therefore ready to embark on the second part: understanding the foundations of how certain models work. However, my research then turned into a constant race to stay up to date, where it was extremely difficult to keep up the pace while pursuing my research and trying to constantly update it with new LLMs. Moreover, without the appropriate technical and financial resources and the growing complexity of the models, it was becoming increasingly difficult to apply these LLMs in their entirety to study their behavior. I therefore decided to let go of LLMs and study these phenomena on simpler architectures, letting me experiment, compare my results with equivalent models, and employ and analyze them in the context of automatic document summarization. Finally, I would say that this thesis is still relevant today, because even if it does not concern directly LLMs, it attempts to tackle a profound task of understanding a dimension of the language that will hopefully shed light on principles that can ultimately be applied to current models.

On a more prosaic note, I wish to add some complementary information about the work that has been done. This document is an article-based thesis, with the involvement from several authors. Therefore, it seems essential to explicit the contribution of each of them for the articles, to reassure readers that this thesis and these articles are the result of my effort. First, all the ideas and formulas developed in these articles originate from my work. Two Masters students took part in my articles, collecting data and coding prototype version of algorithms, in particular for some baselines. During that period I supervised them, and I debugged their code and adapted it afterwards to solve the research questions raised. Two professors, including my PhD supervisor, were involved in writing the articles. Their contribution was limited to complementary suggestions to arrive at a more solid work and a publishable article.

General Introduction

In April 2023, there were 5.18 billion internet users and 4.8 billion social media users.[1] This quantity of people on the web generates an enormous volume of content mainly textual data. To give a few examples, in 2022 there was an online production of 16 million messages, 231 million emails, or 350 thousand tweets per minute.[2] If we look over a more extended period, we can then account for 2 trillion posts shared and, of course, collected on a platform such as Facebook.[3] The wide-scale digitization of classic communication structures allows people and companies to publish more content in different formats and intended for different audiences. Social networks, corporate blogs and web pages, online news media, books, scientific literature, customer opinions, and email communications are all part of the phenomenon of information digitization. The volume and variety of online text data become unmatched by any other source. Moreover, its complementarity with more traditional structured data makes it a precious medium for companies and their analysts to understand their economic environment and consumers better and improve their decision-making (Gentzkow et al., 2019). In the finance and banking industry, we can use the Net Promoter Score, one of the most used indicators, to appreciate a customer's experience and compare institution efficiency (Reichheld, 2006). The score is provided on a scale of 1 to 5, describing whether a client would recommend the enterprise to these friends; it is also composed of a comment which details the reason for the grade. Once analyzed, this knowledge completes standard evaluations by better understanding

[1] https://www.statista.com/statistics/617136/digital-population-worldwide/

[2] https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/

[3] https://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats/

customers' voices and thus improving the products and services offered. However, in this case, as in many others, the sheer quantity of documents available makes it difficult to access relevant information easily. This has increased interest in forms of technology to create overviews so that this textual data can be utilized effectively, such as information retrieval, question answering, and automatic document summarization systems.

Automatic text summarization

Automatic text summarization is the process of distilling the information contained in a single or multiple sources to produce a reduced version of the original material by means of a computer to fulfill a purpose and meet a specific user need (Mani, 2001; Hovy et al., 1999). This process then encompasses a collection of different tasks that encounter these needs. Single long documents, multi-documents, opinion-oriented, aspect-based, or even update summarization are well-known examples of such diversity. The first models were introduced in the late '50s and '60s and aimed to create abstracts of scientific papers in chemistry (Luhn, 1958; Edmundson, 1969). These first models were predicated on a set of heuristics combining statistical and linguistic methods to extract relevant information. Increased involvement in automatic text summarization due to the proliferation of available data on the internet drove the interest in having concrete common resources and structures to evaluate and analyze the different approaches in real contexts. The advent of document summarization conferences such as TIPSTER Text Summarization Evaluation SUMMAC,[4] the Document Understanding Conference (DUC),[5] or the Text Analysis Conferences (TAC)[6] made it possible to make clean and annotated datasets accessible to the community. These conferences have been instrumental in providing a normalized control framework on various tasks, and the datasets continue to be improved, enriched, and employed by current researchers to analyze and compare the performance of their systems. The proliferation of these datasets has, therefore, induced the emergence

[4] https://www-nlpir.nist.gov/related_projects/tipster_summac/

[5] https://duc.nist.gov/

[6] https://tac.nist.gov/about/index.html

of methods grounded in machine learning, and supervised models have become the most studied techniques in the literature in recent years. Extracting the most relevant sentences to include in a summary transitions to a binary classification task, and researchers have trained different types of classifiers to solve this problem (Kupiec et al., 1995; Conroy and O'leary, 2001; Aone et al., 1998). Many machine learning techniques can be employed, and we refer the lecturer to the multiple literature reviews on these approaches (Gupta and Lehal, 2010; Lloret and Palomar, 2012; Ježek and Steinberger, 2008). Following the success of deep learning systems, several methods were also developed for extractive (Kågebäck et al., 2014) or abstractive summarization (Rush et al., 2015). Finally, the breakthrough of using pre-trained Large Language Models (LLMs) based on transformer architectures such as BERT (Devlin et al., 2018), or GPT-2 and GPT-3 (Radford et al., 2019), or T5 (Raffel et al., 2020) have recently allowed to obtain more meaningful representation with prior knowledge and methods as BART (Lewis et al., 2019) or PEGASUS (Zhang et al., 2020) are now the state-of-the-art tools for abstractive summarization.

Whatever method is employed, it is based on a theoretical background from the behavioral study of humans when they must perform such a task. Hidi and Anderson (1986) define summaries in their work as short statements that abridge the information and reflect the gist of the discourse of an original document. The authors also explain that different steps are required to establish a good summary: comprehension, evaluation, condensation, and frequent transformation of ideas. All these steps, depending on the length of the desired output, become a choice about what the most important information in the source document is (Hidi and Anderson, 1986). The way of approaching summarization copies this mechanism and is decomposed into three major steps as detailed by Nenkova and McKeown (2012) in their survey:

1. Comprehension: Systems need to learn a representation of the original texts to fulfill users' needs. The representation will imply, for example, to focus on text representativeness with a graph-based strategy such as in *TextRank* (Mihalcea and Tarau, 2004a) or to stress sentence specificity using TFIDF bag of words vectors

(Ledeneva et al., 2008).

2. Evaluation: Systems must contain a function to evaluate the relevance of text segments to include in the output. Such function can promote centrality by selecting centroids of clusters as representative texts such as in MEAD (Radev et al., 2004). They can further support diversity in sentence scoring with maximal marginal relevance approaches (Carbonell and Goldstein, 1998; Boudin et al., 2008).

3. Condensation: Systems must tackle the summary generation process. Optimization methods can impose sentence relevance while constraining summary length (McDonald, 2007), and enforcing some linguistic features (Ganesan et al., 2010), or directly maximizing token likelihood (Rush et al., 2015).

Factors influencing summarization methods

These three steps, which make it possible to characterize all document summary approaches generally, do not make it possible to distinguish and understand why certain models will perform better than others in specific contexts. Indeed, a system producing a general-purpose summary, even a very good one, cannot respond to all tasks and user needs (Jones et al., 1999). There are, therefore, structural factors linked to the task, the data, and the use of summaries which will influence the functioning of document summary systems, and which can be grouped, once again, under three categories of factors (Hovy et al., 1999):

- Input factors, which represent the characteristics of the input document(s) and how they should be represented:

  - Specificity—specific or general field: The input document(s) can belong to a field, which may have a specific content, compared to a more diverse and general case. For example, news sources often focus on particular events providing answers to unique questions: *who, what, where, when, why* (Owczarzak and Dang, 2011).

  - Genre—Input document(s) can be newspapers, scientific papers, meeting transcripts, opinions, books, and so on. These genres have highly varying formats and grammatical conventions..
  - Source size—single or multiple documents: A single summary contains information from a single document like a book whereas a multiple summary includes the content of a set of document(s) often assumed to be thematically related such as customer opinions on a product.

- Output factors, which represent the characteristics of the generation of the final production:
  - Derivation—extract or abstract: An extractive summary is a collection of segments of the original input whereas an abstractive summary is a newly generated piece of text.
  - Coherence—fluent or non-fluent: A summary can be understable for humans, following the rules of coherent discourse structure, or not.
  - Partiality: The summary can represent the main personal opinions and points of view of the author(s) or it can represent objective and balanced information.

- Context factors or purpose, which refers refer to the relation between input and the output summary and the assessment of relevant information:
  - Audience—generic or specific: A generic summary provides an overview of the input that covers all themes in it while a specific summary focuses on targeted themes that meet a defined user's need.
  - Usage—indicative or informative: an informative summary reflects what the source says about something and describes it while an indicative one allows the user to understand the topic of the input without knowing its full content.
  - Situation or Task: It refers to the context within which the summary is to be used (who by, what for, and when). A summary for describing customer opinion for a company is different as an abstract for a board of review.

All these factors influence the creation of summarization algorithms since they will modify each stage described previously. To give simple examples of the impact, the representation of texts depends on input factors: a system based on the structure of the discourse to embody a text (Marcu, 1997) cannot be applied to characterize multiple short tweets. The evaluation function will be affected by the user's needs. Textrank (Mihalcea and Tarau, 2004a) depicts the central topic to inform about the subject mentioned, while a model founded on topic modeling (Gong and Liu, 2001) indicates which theme is discussed in the documents. Finally, generation can be easily influenced by these factors since we can try to maximize the material coverage via the diversity of the generated sequences or selected sentences.

Humans produce better summaries after being trained to identify relevant source texts such textual features as topic sentences, keywords, and repeated ideas (Hidi and Anderson, 1986). Therefore, the same phenomenon is expected to appear in the summaries used by the automatic text summarization community, especially for supervised learning. Indeed, in this case, the definition of the relevance of the information and purpose becomes underlying since the algorithm's operation and, the text's characterization and the information's importance is done through the labels provided to the model. However, when no instructions are specified, the human summaries used, although different, are based on shared properties such as the use of term frequency (Nenkova and Vanderwende, 2005), including named entities, subject-specific terms, and the non-inclusion of reported facts and figures (Goldstein et al., 1999). Although many distinct instructions and tasks have been proposed at various conferences such as DUC, NIST, TAC, or other datasets, authors observe that the purpose or intention of the summary is never stipulated in these tasks (Over et al., 2007) and that the tasks are now always specific since the community was not satisfied with generic summaries. So there is no reason to believe that, regarding these objectives, the human experts producing those outputs follow their natural tendencies, especially when we know that the most used data in the literature are news stories that tend to employ events and named entities (Filatova and Hatzivassiloglou, 2004). This phenomenon thus establishes an implicit homogenization of summaries format toward specific indicative texts (Jones, 1972, 2007). This is consistent with well-known findings on the issues related to using human gold standards since it has been demonstrated that, first, human versions still have significant variance in their output (Voorhees and Tice, 2000) depending on the tacit perception of the purpose of the abstract by the annotators (Sjöbergh, 2007). Second, another major problem due to the existence of all these various tasks, datasets, and purposes is that these supervised approaches lack portability and reproducibility (Mihalcea and Tarau, 2004b), creating a monstrous need for a workforce to train enough models to perform on all these tasks. Finally, the metrics associated with supervised learning also suffer from bias toward lexical similarity and do not account for fluency and readability (Scialom et al., 2019), or that they are easy to fool and that one can obtain outstanding scores without producing a good summary because they rely highly on a frequency count where greedy methods can achieve better than a consensus of human experts (Sjöbergh, 2007).

Unsupervised summarization

Unsupervised methods have always been favoured for document summarization, whether for design flexibility, access to data, portability, or evaluation difficulty. Still, they are today with the advent of deep learning and LLMs (Khosravani and Trabelsi, 2023). Moreover, we must remember that the objective of the summary is to make it easier for humans to digest information; the very idea of data labelling seems contradictory. The advantage of unsupervised models is that they offer researchers and designers the full possibility to choose the encoding of documents, the content evaluation, and the text generation. In other words, complete control over the 3 steps of creating a summary system. It also means that researchers must consider all the factors influencing the summary to implement an effective method. The impact of input factors, such as data domain or specificity, and output factors, such as summary derivation, are more obvious to characterize and therefore better studied in the literature (Gupta and Lehal, 2010; Lloret and Palomar, 2012; Ježek and Steinberger, 2008). However, they remain superficial factors that may lack clarity to convey the underlying phenomena that could explain discrepancies between similar techniques or results between various techniques. In particular, to complete different tasks, the writer of the summary must create relations between segments in the text and relate them to their personal knowledge base and experience (Wittrock and Alesandrini, 1990). Recognizing that, in most current summarization cases, the target is almost always another person, the relation with the recipient's experience and knowledge should affect the generation process (Hill, 1991). These individual attachments thus establish different abstract and unconscious intentions and purposes in constructing the summary, impacting the perception of what significant information is (Belkin et al., 1982). In the case of unsupervised models, the underlying structure of the data is used to produce an intermediate representation that links the input document and the summary. This intermediate structure is depicted by the diverse states that characterize the notion of information relevance (Lavrenko, 2008). These general notions of information delineation will, therefore, also be influenced by the user's induced state of knowledge and their intention, and they will profoundly affect the functioning of the model and its performance (Peyrard, 2018). Moreover, as it has been presented in Jung et al. (2019), many automatic summarization methods can have biases toward certain information, and the authors show that this tendency is especially true for unsupervised methods. Although it is easy to see the purpose factors emerging by reading papers, it is unfortunately never explicitly defined for most document summary approaches. It ultimately poses problems since models which are not necessarily designed to meet the same information needs are employed similarly. Of course, this further restricts the analysis of these models and limits understanding why some models perform better than others on certain datasets or tasks, especially when we compare them based on human references where the intention was again not specified.

Thesis structure and contributions

Since many automatic summarization methods can have biases toward certain information, especially unsupervised methods (Jung et al., 2019), this thesis intends to focus on unsupervised methods applied to document summarization and their connection to purpose factors. In this work, we do not aim to propose one absolute definition of relevance but rather spotlight existing differences and adopt them as a new way to highlight the field's historical approaches and recent techniques to understand how they are related, particularly how the models have evolved with deep learning, and how they compare to each other. By taking this angle of analysis, this thesis frees itself from models' superficial characteristics to accentuate how they differ in the way they encode, evaluate and generate content. It will demonstrate how to employ these concepts to bias methods to answer better specific users' needs or various tasks. Our first contribution to the current literature is to provide this analytical framework with a new typology, establishing a link between topic relevance and information coverage with the notions of salience, representativeness, redundancy, and diversity. It also contributes by emphasizing how these different characterizing elements relate to common attributes of unsupervised summarization methods, creating a bridge between information and summarization stages. Our third contribution is an empirical demonstration of the importance of taking these aspects into account. Indeed, we present three concrete examples where content has been biased to improve those of generalist algorithms on particular tasks. Our first model will evidence how to play with the concept of redundancy to control the cohesion or novelty in the representation and evaluation of information for update summarization. Our second model will illustrate how to modify the evaluation function of importance to add objectivity to an opinion summary. Finally, our last approach will demonstrate how to exploit topic modeling to increase diversity in summary generation. In the discussion, we aim to investigate the link between our work, the traditional stages of summarization, and information bias. We will emphasize the new evaluation methods proposed in our papers, which have let us highlight the contribution of these methods. Indeed, our last contribution is to provide a discussion on the issues with the data sets and performance measures of unsupervised models since it is not intuitive to imagine the utilization of summaries created by humans as a reference for evaluating systems whose purpose is not to rely on such information. Finally, we will use all our analyses to suggest new ways of thinking about the different characteristics of information and how they can participate in unsupervised document summarization research and the assessment of large language models.

The remainder of this thesis will be structured in the following manner.

First, we introduce our literature review for unsupervised summarization approaches in our section **Theoretical framework** . This literature review presents different facets of how relevant information can be encoded. We propose distinguishing between selective and central information to characterize important content for a summary. We also introduce two facets of content coverage. We separate methods that promote information gain by reducing redundancy with the ones that support diversity. The main advantage of this new typology is that it proposes a non-superficial classification, which remains consistent with both historical and recent approaches for document summarization and facilitates their comparison by highlighting profound similarities. Besides, linking contextual purpose factors and relevance puts the users' needs and usage back at the heart of the methods. In the remainder of this thesis, we exploit the idea that distinct notions of relevance can meet different needs through three concrete applications.

Chapter 1 **Unsupervised update summarization of news events** introduces our first article published in *Pattern Recongition*. The paper presents the concept of unsupervised update extreme summarization and proposes a novel competing architecture between information and a language model to handle the task. The model relies on a combination of TFIDF, reconstruction constraints, and an update parameter to identify relevant and consistent material to preserve in the summary. The model was tested on a news dataset and performed strongly in this sentence compression context. A master's student implemented the autoencoder for the language model and the baselines, and the third contributor to the paper provided advice and correction.

Chapter 2 **Objective and neutral summarization of customer reviews** introduces our second paper. In this work, we initiate the novel task of objective opinion summarization along with an unsupervised model to do this task. Specifically, we modify a general autoencoder architecture with gradient reversal layers to learn sentiment-agnostic sentences. Relevant information is thus represented by a central average and objective content. We also created a new dataset based on product reviews for objective summarization and several automatic metrics for evaluating objectivity. A master's student did the baseline autoencoder and the data collection. The two professors provided recommendations on the experimental protocol and the paper writing.

Chapter 3 **Topically diversified summarization of customer reviews** introduces our last paper presented at the ICNLSP 2023 conference. The paper proposes a variational autoencoder architecture combined with a topic model approach in a multitask learning framework. This approach allows the system to generate user-oriented summaries where relevant information is identified by its belonging to the main topic defined by the user. It also produces generic summaries by diversifying the number of topics addressed in the summary. The model was evaluated on product review summarization, and several metrics to assess diversity and input coverage were introduced. The professor and co-author of the paper provided recommendations on the design of evaluation metrics and the experimental protocol.

Finally, in the **General Conclusion and Discussion** section, we emphasize the contributions of our work. More specifically, we demonstrate how, in those papers, the three main of the summarization process were modified to account for specific tasks. We mainly highlight how the representation, the scoring, and the generation of information can be biased to address different usages and audiences. This then lets us link with our literature review and generalize our observations. This leads us to discuss the connection between purpose factors, such as indicativity, informativeness, specificity, and genericity and their relationship to the relevance encoding. We will discuss the implications of clarifying this link and make suggestions for further development of our framework.

Theoretical framework

A summary is a short statement that abridges the information and reflects the gist of the discourse of an original document. The task becomes a choice about determining the most important content in the source document (Hidi and Anderson, 1986). In most of the cases, the summary is meant to be used by another person, therefore the relation with the recipient's experience and knowledge affect the generation process (Hill, 1991). These individual attachments thus establish distinct abstract and unconscious intentions and purposes in the construction of the summary, impacting the perception of the significant information. This is called an abnormal state of knowledge (Belkin et al., 1982) and it influences the interpretation of important material in the documents. This intermediate structure is depicted by two concepts that characterize relevance: topical relevance, which identifies the information to be presented to the user; and novelty relevance, which considers the contribution of knowledge made to the user (Lavrenko, 2008). As figure 0.1 shows, this difference in perceived importance influences the various stages in creating the summary.

The aim of proposing such a literature review is, of course, to bestow the framework in which the papers written are included. But it is also to propose a new typology that will let us understand why the real contributions of these papers. Indeed, the models submitted are often based on a general architecture that will be modified to meet a specific need. During this literature review, we will further introduce the set of metrics usually employed to measure unsupervised approaches' performance. This will allow us not only to present their initial limitations but also to understand why new ones have been created uniquely

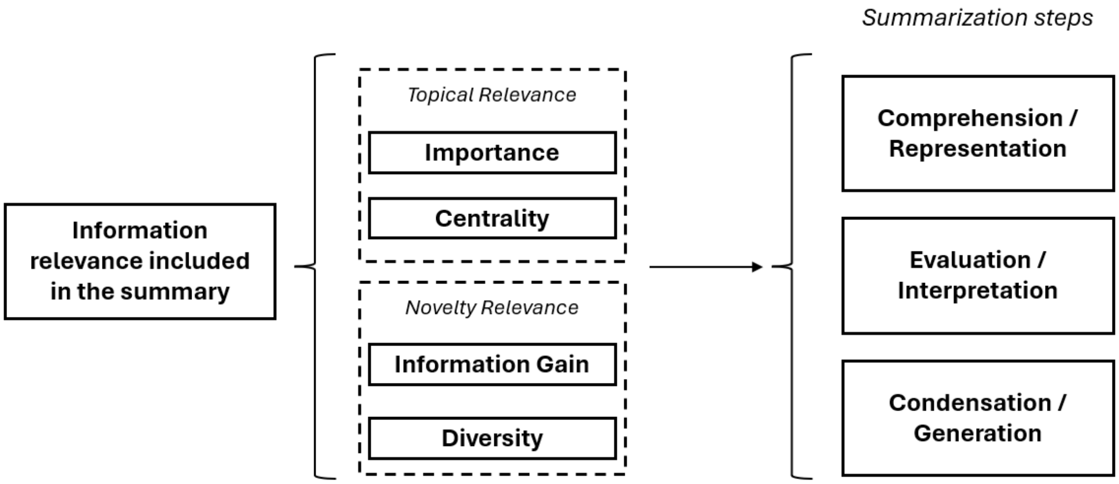

Summarization steps
Information relevance included in the summary
Topical Relevance
Importance
Centrality
Novelty Relevance
Information Gain
Diversity
Comprehension / Representation
Evaluation / Interpretation
Condensation / Generation

Figure 0.1: Information relevance can be represented in several dimensions. The different stages of summary creation will be impacted accordingly.

for the thesis papers. This section will therefore be organized into 3 sections. We'll start by defining topical relevance, classifying methods between two notions: salience and centrality. We will do a similar work for the notions of redundancy and coverage in novelty relevance. Finally, we present the intrinsic and extrinsic metrics related to unsupervised methods for document summarization.

Topical relevance

Since Luhn (1958), unsupervised models heavily depend on the notion of relevance to design the function that will evaluate importance scores to textual units. In their work, Peyrard (2018) extends and generalizes the definition of relevance as the measure that minimizes information loss between the text and its approximation. The author relates this principle to entropy, and the variation between models is principally due to different notions of topical frequency. This interpretation of topical relevance is perfectly suitable for placing this concept, as employed in automatic summarization, into the broader context of information theory. But this definition only considers the relevance to the subject of the source document, and it omits usage factors. To give a simple example, if the requirement is to summarize the main news event, we'll observe the application of the frequency of grammatical or thematic attributes, which will be characterized by terms such as *salience* (Lloret et al., 2008; Hatzivassiloglou et al., 2001) . While notions of *centrality* with the use of features similarity will be found (Radev et al., 2001; Erkan and Radev, 2004) when the goal is to integrate the news summary into a search engine and thus provide a global view for text retrieval. These notable differences lead us to divide topic relevance into two categories related to salience and centrality.

Salience: relevance as selective information

To understand why researchers make choices for representation, evaluation, and generation, it is essential to specify a formal statement of what we call an *ideal summary*, which defines the objective that unsupervised models try to reach. Humans, when asked to summarize, tend to produce a condensed version of the text, containing the most important information in it (Banerjee et al., 2015). Consequently, they capture key characteristics or events related to what they perceive as the major content in the original text (Maybury, 1995). Trying to reproduce this behavior, automatic text summarization systems attempt to state the meaning of this main content by targeting the part of the input texts that are relevant to the main topics (Barzilay and Elhadad, 1999; Jiang et al., 2018). The goal is then to assess the relevance of those terms to identify and select the one to preserve (Steinberger and Jezek, 2004). An ideal summary is an approximation containing the most important topics of the original document. These main topics are related to different specific and characteristic concepts, events, and aspects (how, when, why, etc.) that should be included in the summary. Relevance is therefore based on this perception of targeted or selective relevance and becomes the way to discriminate which elements will be characteristic of those concepts.

The problem here consists in determining which are the main topics and thus which information and is worthy of inclusion in the final production (Nenkova and Vanderwende, 2005). Following this principle, multiple methods arise to score the importance of such information. The first techniques observe the assumptions of (Luhn, 1958), where the word frequency characterizes importance. Further legitimized by the work of Nenkova and Vanderwende (2005), who prove that humans focus on frequency to produce their summaries, multiple methods have been implemented along these lines. There are approaches that directly follow the frequency (normalized or not) of terms such as *n*-grams (Gillick et al., 2008; Gillick and Favre, 2009), conceptual units representing events (Filatova and Hatzivassiloglou, 2004; Takamura and Okumura, 2009), or general keyphrases (Riedhammer et al., 2008, 2010). Once these textual units are defined, a first approach can measure the capacity of a model to detect these important unit from a generated summary by a large language model (Fu et al., 2021). More traditional approaches rely on the objective to maximize their presence in the final summary. The well known *Term Frequency Inverse Document Frequency*, or *TFIDF*, metric is used to emphasize the specificity of a term and has demonstrated solid performances in numerous language processing tasks; it is thus logical to see that several papers use this metric to determine important terms (Nomoto and Matsumoto, 2001a, 2003; Amini and Gallinari, 2001; McDonald, 2007; Christian et al., 2016). In the context of the new deep learning techniques, TFIDF is used to select relevant terms or to mask them to create a constraint for a language model to include these terms in the summary (Laban et al., 2021; Carichon et al., 2023). In a similar vein, Févry and Phang (2018) consider grammatical words to be unimportant. They therefore increased the sentence size with unimportant terms and used a denoising autoencoder architecture to learn a language model filtering the presence of these terms. In addition to this technique, mutual information (Ganesan et al., 2012), information gain, and residual inverse document frequency are also used to score importance (Lloret and Palomar, 2012). We once again note the same idea for deep learning models to create new constraints for a language model this time depending on improbable informative words (Malireddy et al., 2020), mutual information between the original sentence and the summary (West et al., 2019), or weighted-pooling operation on attention weights paid to keywords (Zhang et al., 2023a). Finally, some new authors have characterized term importance directly using ROUGE score (Lin, 2004) from the source documents and pseudo summaries. Once the terms are identified they can directly be used to filter information for fine-tuning large language models (Bražinskas et al., 2020). This framework displayed strong performances to improve the utilization of large language models to generate summaries in a zero-shot learning context (Rothe et al., 2021; Fabbri et al., 2020). Another frequency-based hypothesis for isolating the contribution of terms in a unigram or multimodal language model (Vanderwende et al., 2007; Nenkova and Vanderwende, 2005). Specifically, we measure how probable a term *t* is, given a trained probability distribution model on a background corpus (Conroy et al., 2006; Zopf et al., 2016). Because of their properties, probability distributions are a good way to create a language model and thus combine frequency with sentence structures. Graph-based representations, especially directed by co-occurring networks, are a good means to represent this sequential structure. By creating proper metrics for weighting edges, such as frequencies or transition probabilities, the graph structure can be used for text summarization (Filippova, 2010). Indeed, the importance of topics can be depicted by identifying recurrent term sequences, as represented by the shortest paths between significant lattices (or predefined starting nodes) (Ganesan et al., 2010; Cardenas et al., 2021; Shang et al., 2018).

Early on, researchers thought of enhancing and enriching these statistics-based features with other specific, data, or task-related elements. The first we can cite is the use of structural information. The hypothesis is that sentences located at strategic positions in a document may contain more important topics. First works include a scoring method based on the absolute location (Edmundson, 1969), but relative position scoring has been privileged later for its better performance (Ferreira et al., 2013; Oliveira et al., 2016), and especially the average position of terms in the text (Yih et al., 2007). Another inference concerns scoring sentence length, because it is supposed that succinct sentences are irrelevant in terms of topic representation, while very long ones are a waste of space in a summary (Ferreira et al., 2013; Oliveira et al., 2016). Then, additional approaches consist in using external knowledge to assess the importance of terms. The first approach to apply this assumption introduced cue bonuses and stigma words to reward or penalize sentences depending on the presence of these terms (Edmundson, 1969). Some later researchers further specialized these cue expressions to specific fields, for example Arab politics (Al-Radaideh and Bataineh, 2018). Finally, some authors have used general knowledge bases such as Wikipedia to enrich the semantic information provided by some terms (Sankarasubramaniam et al., 2014). The features added here directly concern statistical or general structure features, but further specialized features have been employed depending on the input data or the task, such as the overlap between titles or headings for government reports (Edmundson, 1969), the use of numbers or dates (Schiffman et al., 2002; Oliveira et al., 2016) or named entity (Xiao et al., 2021) for newspapers, citations of other researchers for scientific papers (Abu-Jbara and Radev, 2011), sentiment polarity and subjectivity for opinions (Anuradha and Varma, 2016), or even similarity to a query in query-focused summarization (Jin et al., 2010). This specialization can be used with deep learning techniques, such as autoencoders, to automatically learn abstract features based to overlaid onto the task. For example, some authors proposed to use sentiments as labels for opinion summarization (Denil et al., 2014). Other authors also observed that traditional methods did not work optimally when summarizing opinions (Tampe et al., 2022). Therefore, they completed their autoencoder model with an attention system weighted by the number of likes of the tweets, thus allowing completing estimation of term importance with the information of popularity. Finally, for opinion summarization, Amplayo et al. (2021) masked predefined seeded terms representing aspects of a product to fine-tune a large language models and compel the model to include them in the generated summary. All these new enriched representations thereby improve the capacity of models to capture significant topics.

Another explored possibility is to design models that directly link terms and topics. The first attempt was made through the use of *topic signatures* (Lin and Hovy, 2000). In this framework, topics are represented by a concept, often predefined, and signatures are terms that are highly correlated with the concept. Some authors suggest ways to enhance the term-to-topic association by stating that documents belonging to the same topics have the same probabilistic content models (Barzilay and Lee, 2004). Others propose rather to generalize the concepts to themes affirming that topics are related to each other (Harabagiu and Lacatusu, 2005). Once themes are pinpointed, they must be scored for importance for summarization, using sentences most tied to the topics and then they must be extracted. Unfortunately, these methods count on supervision to identify topics and still do not consider interrelations between words. The first step towards designing these associations and to unsupervised topic modeling is to apply dimensionality reduction approaches to a bag-of-words model. Methods as *Singular Value Decomposition* (SVD) capture recurring semantic relationships between terms, and these patterns directly characterize topics in documents (Gong and Liu, 2001). The more pattern occurs in a document, the higher its singular value. Thus, choosing the sentences most related to the k-best values is a first technique to select the most important topics (Gong and Liu, 2001). *Non-negative Matrix Factorization* (NMF) improves SVD approaches by constructing nonnegative part-based representations constraining to have positive values in the topic matrices and that is more natural for textual interpretation (Tsarev et al., 2011). Furthermore, vectors obtained with NMF are sparser than with SVD, granting a better association between topics and sentences (Lee et al., 2009). *Latent Dirichlet Allocation* (LDA) is a generative method for topic modeling, allowing flexibility on the hyperprior distribution of terms. Once LDA is performed, the significance of topics and terms/sentences can either be directly determined with the estimated probabilities, admitting that α values of the Dirichlet distribution represent the absolute importance of topics (Arora and Ravindran, 2008; Wang et al., 2009), or by a creating bipartite graph structure with topics and sentences and the use of importance calculation algorithms such as PageRank or HITS (Parveen et al., 2015). As for deep learning, some authors have shown that the application of autoencoders makes it possible to create concept-oriented vectors (Yousefi-Azar and Hamey, 2017). These vectors mimic the characteristics of topic models and identify key patterns of terms to include in a summary. Once these representations are learned, the features can be used to determine the importance of text segments by once again summing them or by computing weighted representations (Singh et al., 2016). The use of specific deep learning methods such as denoising techniques also allows isolating some concepts that are considered more important topical aspects (Amplayo and Lapata, 2020). Once the most important concepts are identified, optimization framework is implemented to maximize their presence into our constraint length summary.

Finally, textual cohesion states that salient topics will be discussed throughout the input text, with semantic relations linking all the terms connected to that topic (Gupta et al., 2011). The use of lexical chains is thus a natural choice, given that they represent lexical cohesion relations as categories and pointers to the original document (Barzilay and Elhadad, 1999). To generate a summary, we once again need to define which lexical chains will be most important. Multiple methods have been proposed to identify these strong chains. The length and the homogeneity of a chain can be a strong marker (Barzilay and Elhadad, 1999), but so can the relations between the members of the chain (Brunn et al., 2001; Doran et al., 2004), or their relative frequencies (Gupta et al., 2011). Finally, what remains is to select the above-average ones to take the best ones. Richer representations can be used to improve identification of the central topics. By relying on the coherence principle, other syntactic markers such as ellipses, conjunctions, and substitution references can bring complementary structural information to lexical cohesion (Lynn et al., 2018). A coherent text follows a specific discourse structure, and *Rhetorical Structure Trees* (RST) are objects meant to describe these relations between segments of text (Ono et al., 1994). They make it possible to distinguish important structural elements (nuclei) from weak ones (satellites). The objective is then to select important segments or discard lesser ones. Several penalty schemes have been proposed such as the number of connections between nuclei and satellites (Ono et al., 1994), the nature of their connections (O'Donnell, 1997), or promotion sets to characterize relations between nuclei and satellites (Marcu, 1997). Other improvements have been proposed over time and we refer the reader to the dedicated review of these models made by (Uzêda et al., 2010). Other graph models close to RSTs, such as discourse graphs (Christensen et al., 2013) or *Abstract Meaning Representation* (AMR) graphs (Dohare et al., 2017), have also been used because of their accurate representation of intersentential coherent relations. The semantic and syntactic properties of RST or AMR graphs are particularly useful to provide meaningful initial layers for autoencoder systems. Important information is filtered from the graph by using heuristics (Dohare et al., 2017) or domain-knowledge ranking systems (Hou and Lu, 2020) and then the decoder generate a length constraint text to form the final summary. Rather than directly identifying important segments of text, RST structures have also been used to cluster segments by their roles and then use statistical features to select important ones (Atkinson and Munoz, 2013). Some authors have used others grammatical and structural information such as verbal and noun phrases with statistical features to spot important events or aspects and maximize their presence in the summary (Bing et al., 2015). Finally, dependency trees have been used, especially because these structures show interesting properties for the abstractive generation of summaries (Banerjee et al., 2015; Cheung and Penn, 2014).

Centrality: relevance as representative information

Providing information on a few main topics is very useful for understanding the document, but it cannot replace the entire document. For certain needs or tasks, such as indexing in a search engine, it's essential to have a summary that fully portrays the complete document (Radev et al., 2004a). Thus, the ideal summary depicts the best as possible the input document(s) by covering and describing its various themes. The relevant information can be seen as a representation that conveys information as a whole (Takamura and Okumura, 2009) by being the most redundant with other text segments (Radev et al., 2004b), and that enforces a correlation of the semantic volume between the summary and the initial text(s) (Yogatama et al., 2015). The relevance can then be stated in terms of centrality or representativeness, which express the extent of content provided in the original input that is included in the summary (Huang et al., 2010).

To include the core information of the document(s) in our summary, we need to identify the most representative elements of this input. This goal is achievable by using notions as pairwise similarity between text segments to recognize which one are the most like all others. Clustering groups together data with analogous properties. Once these elements have been grouped together, we can identify a centroid for this cluster. This point subsequently represents the barycenter of the information contained in our input segments.

The MEAD system was the first model that applied centroid as a pseudo-document with terms features above a certain threshold to symbolize the center of the segments (Radev et al., 2004b). We can then employ the text segments that are the closest to this centroid to form the summary (García-Hernández et al., 2008; Song et al., 2011). Multiple variations have been made to this method, either by changing the representation of the document(s) for richer ones such as word embeddings (Rossiello et al., 2017; Padmakumar and Saran, 2016; Lamsiyah et al., 2021b), by testing different clustering algorithms to use optimized fuzzy evolutionary algorithms (Alguliev et al., 2009; Song et al., 2011), as well as considering word-level features then grouping them to cluster interesting segments (Banerjee et al., 2016; Ferreira et al., 2014), or hard singular value decomposition and maximizing proximity to the centroid through a greedy method to increase the semantic volume (Yogatama et al., 2015). Some have considered adding semantic features such as WordNet information (Huang et al., 2010) or even full syntactic data, to evaluate paraphrasing and detect themes (Barzilay et al., 1999). Once the cluster have been created, they can be used as input for abstractive methods by providing only central sentences to deep learning models (Xiao et al., 2021). Coavoux et al. (2019) creates clusters of opinion and feed the review constituting the main cluster to an autoencoder architecture to generate the summary. These clusters have also been employed as input to pre-trained or fine-tuned language models to condition text generation (Suhara et al., 2020). Finally, Angelidis et al. (2021) proposed an approach where clusters of latent representation are learned dynamically during training process to select central representations and produce general summaries of customer opinions.

This notion estimates proximity the most representative point does not use directly pairwise similarity and can present some flaws, such as not considering sentence subsumption and being too sensitive to rare words (Erkan and Radev, 2004). Therefore, some approaches have proposed to directly evaluate the closeness of each segment to every other (Ribeiro and de Matos, 2011). Because of their properties, especially for carrying global text information, most of the methods use affinity graphs, such as the kNN similarity graphs or the ε-graphs. In these structures the centrality is inspired from the prestige concept in social networks (Erkan and Radev, 2004) because each link between vertices can be casting a vote or recommendation for those nodes (Mihalcea and Tarau, 2004). Once the graph is constructed, there are many approaches to rank and select the most similar sentences. The LexRank algorithm (Erkan and Radev, 2004) employs random walks that would determine the most probable node of the graph and was improved to account for Markov chains hypotheses in these walks for the Grasshoper method (Zhu et al., 2007), and the CoreRank system (Fang et al., 2017). Mihalcea and Tarau (2004) perform several modifications in their work by adding oriented edges to the graph and by changing the approach by considering that important nodes give stronger recommendations to its peers than weaker ones. This popularity-based method is reflected using adapted versions of the well-known PageRank and HITS algorithms. Finally, some authors employed the idea by modifying the selection of sentences exploiting shortest path estimation in the similarity graph (Thakkar et al., 2010). Another theory assumes that nearby points are likely to have the same ranking scores, thereby making the manifold ranking technique appropriate to perform node selection (Wan and Xiao, 2009). These methods are being extremely efficient, several authors have proposed modifications to them. The first adjustments introduce special data, such as citations for scientific papers (Qazvinian et al., 2013), special external resources and lexical features (Leite et al., 2007; Heu et al., 2015), or structural metadata, by reinforcing links with intra-document information (Wei et al., 2008, 2010), or with the hirearchical structure of the document (Dong et al., 2020). Additional improvements create richer sentence representations with deep learning techniques to strengthen similarity estimation (Yin and Pei, 2015; Alami et al., 2018; Zheng and Lapata, 2019). Once again, a very recent approach transform this technique for abstractive summarization by selecting representative sentences with similarity graphs and input them to a pretrained large language models to form more coherent outputs (Zhang et al., 2023b). Other authors have proposed drastically different graph techniques to measure sentence centrality, such as using InfoMap and clustering coefficients (Dutta et al., 2019), or by employing graph cuts to select subsets representing the summary then maximizing pairwise similarity through submodular optimization (Lin et al., 2009; Lin and Bilmes, 2010; Kågebäck et al., 2014).

Finally, some researchers take a more direct approach by dynamically optimizing the pairwise similarity between passages of texts (McDonald, 2007). While the authors estimate the pairwise similarity via the cosine distance and generate the summary by integer linear programming, some methods are based on intersections of hyper-planes formed by the sentences in the word space (Vanetik et al., 2020); or some have measured the coverage according to the capacity of a sentence to reconstruct the other sentences (Liu et al., 2015), and the construction of the summary is optimized by an algorithm of simulated annealing making it possible to take into account the sparsity issues. These approaches can be further specialized to avoid noise by dealing with passages on the same topics by using either clustering techniques (Alguliev et al., 2009; Song et al., 2011), topic signatures (Dias and Alves, 2005), or by constituting homogeneous item sets of related sentences (Ribeiro and de Matos, 2011).

However, in this framework the relation between the input and the output is not explicit. One way to formalize this problem explicitly is to minimize the reconstruction error based on similarity when selecting the segments. The first methods directly depict each element/sentence with vector representations then create optimization procedures based on L2 and L1 constraints on the different elements and on a selection matrix to force the selection of the minimum number of closest segments to select the best ones to include in the summary. Some authors have used (Yao et al., 2015). Others have used *n*-grams or embedding representations combined with Kullback-Leibler divergence to minimize the difference in the probability distribution of these elements (Peyrard and Eckle-Kohler, 2016; Kobayashi et al., 2015). Instead of trying to optimize the reconstruction for each segment, some authors have highlighted the interest of having an average representation of the whole input to capture its overall content (Ma et al., 2016). New deep learning approaches are another solution to enhance row features since they can model nonlinear relations between terms, creating a better approximation of the human cortex's way of functioning (Liu et al., 2012). Thus, the authors have employed a paragraph vector model (Le and Mikolov, 2014) to create this mean, then use Euclidean distance to minimize the difference between summary and document. These average features can also be used to determine the importance of text segments either by once again summing them (Singh et al., 2016) or by using the new importance vectors in optimization processes (Liu et al., 2012; Zhong et al., 2015). Restricted Boltzmann Machines trained with entropy or Kullback-Leibler (KL) are very well suited to enhance the properties of feature matrices and create complex abstract representations. When provided with feature vectors and trained, the algorithms identify the most important terms/features for reconstructing this input. The implementation of document embedding can also be exploited to mark the importance of a segment in the reconstruction error. Some approaches build the whole document embedding, then rebuild it by removing one segment at a time (Joshi et al., 2019). If the distance between the two vectors is significant, it means that the segment is key to capturing the document's overall content. Other methods adopt this principle by estimating the semantic similarity between embeddings of the sentences in the summary and the documents (Schumann et al., 2020). These embeddings are generated using the average representation of the word vectors. Once these representations are created, the authors minimize the cosine distance between the original documents and the summary to select the sentences (Schumann et al., 2020; Liu et al., 2022) while adding different constraints. Other authors have created an average aspect-based representation of a set of reviews and maximizes the KL divergence between the summary and this pseudo-typical document (Chowdhury et al., 2022). In the case of abstractive summarization, the average is used in the loss function of an autoencoder. Some authors have employed a variational autoencoder model to reconstruct the input while applying a size constraint (Schumann, 2018), thus ensuring to include these most important topical text segments. The information constraint then can be replaced by other objective functions such as respecting the topic distribution of input documents (Baziotis et al., 2019). For multidocument summarization, once again, it is possible to adapt these mechanisms of averaged input representations and to embed them directly in deep learning and autoencoder algorithms. The *MeanSum* method (Chu and Liu, 2019) has proposed a model composed of two main components: an encoder learning the representation of each text, and a constraint system building the average of the representation to reconstruct a summary as similar as possible to the set of initial documents. The system learns to select the central information in the input, thus reproducing the whole original material. Another avenue consists in exploiting the capabilities of variational autoencoders to learn a latent representation of a set of documents to reconstruct iteratively every input hence capturing the core content of these contextual documents (Bražinskas et al., 2019). When the model generates a summary, it then builds an average representation of this information. Following the same process, other authors have used this representation method to isolate the salient features of a set of background text and have designed a variational autoencoder that produces summaries that specifically highlight updated information in dialogue conversations (Zhang et al., 2021). Finally, in for recent approaches, the average representation can also be used as an input for pretrained large language models (Oved and Levy, 2021).

Conclusion

With this classification differentiating salience and centrality, we are now able to understand behavioral variations between seemingly similar methods. As a first example, we can study the topic model-based methods proposed by Steinberger and Jezek (2004) and Gong and Liu (2001). Both approaches rely on bag-of-word model with singular value decomposition to identify topics and include them in the summary. However, one will select sentences containing the most topics, while the other will extract sentences related to the main topic. The sentences selected will be distinct, and will not have the same purpose. This typology also allows understanding why certain methods applied to the identical dataset meet different needs. If we take the case of opinion summarization, the model introduced by Chu and Liu (2019) aims to represent a consensus between opinions, whereas the model proposed by Amplayo and Lapata (2020) focuses on expressing the primary aspects described. Therefore, we have the grasp that central information attempts to depict a general view that will enable portraying the input fully, thus being informative (Erkan and Radev, 2004) while salient information allows extracting key information that can be employed as an indicative summary (Hatzivassiloglou et al., 2001). The bridge between the purpose factors, information, and the 3 stages of the summarization process then becomes obvious and perfectly explains these differences and make it possible to appreciate better why certain models perform so well for specific tasks.

Novelty relevance

In the first sections of this chapter, we have explored topical relevance in their work (Allan et al., 1999), meaning that we have sought, through the various definitions of relevance, to fulfill the user's need related to the intended usage of this information. Even if it is clear now that a summary should provide important information as much as possible (Peyrard, 2018), its usefulness can also be influenced by previously seen material. This new paradigm that considers the user's prior knowledge introduces a notion described as novelty relevance (Lavrenko, 2008) and whose purpose is to meet the broader user's information needs. In a context of limited size, it becomes crucial to ensure the usefulness and thus the novelty of the elements incorporated in the text. Given this importance in assessing a summary's quality, it is normal that the subject has been largely tackled in unsupervised text summarization. Therefore, systems require to evaluate segments in terms of both relevance and novelty to obtain optimal outputs (Zhai et al., 2015). However, when we address novelty, we are talking about two notions: novelty which favors the exploration of a space of knowledge, and diversity that lets us expand that space (Shah et al., 2003). This distinction has already been examined often, especially in information retrieval, where novelty is defined by the necessity to avoid redundancy to find more information on a specific subject, while diversity is defined by the need to resolve a quest for new and various information (Clarke et al., 2008). Once again, we propose to highlight disparities between approaches for managing novelty and adopt them as an original way to characterize information unsupervised text summarization methods.

Information Gain: Novelty as non-redundancy

Aiming at information novelty is crucial in text summarization, especially in this length-constrained environment where repeated content will increase the noise and thus notably degrade the perceived quality of the summary (Lloret and Palomar, 2013). The task of seeking novelty can therefore be seen as avoiding including concepts if they are already related to the output. This principle is known in text summarization as reducing the *redundancy* (Carbonell and Goldstein, 1998). Thus, an ideal summary is a text that includes important or central content from the original document(s) that brings new information to the user. We thus consider that information novelty is depicted by redundancy between variables of the result, and that content will be new if the summary's content gains a new value by adding some information that is nonredundant with information already included to the user (Carbonell and Goldstein, 1998).

One of the first hypotheses to handle redundancy is to consider the human way of dealing with redundancy by assessing directly if a text segment is too similar to elements already selected in the final output. The simplest idea for measuring similarity and redundancy is to assess lexical repetitions and words overlap. These shallow metrics of direct string matching have shown promising results when the redundancy is not too strong in a corpus (Schiffman et al., 2002). One obvious flaw of this approach is that it only considers one word at a time; thus, it is not surprising to see researchers improving it with *n*-gram based similarity measures (Saggion and Gaizauskas, 2004; Tohalino and Amancio, 2018). Aside from directly checking for textual unit overlap, other count-based similarity measures have proved their efficiency for several natural language processing tasks and, naturally, have been used in various papers. We can see the use of more traditional methods such as Jaccard or cosine similarities (Ganesan et al., 2010; Tohalino and Amancio, 2018; Joshi et al., 2019), or more evolved ones such as the use of mixture models (Zhang et al., 2002), or combinations of Jensen and Kullback-Leibler divergences (Toutanova et al., 2007). These similarity measures remain on lexical information and can perform less well than methods also relying on syntactic and semantic content (Lloret and Palomar, 2013). One good way to improve the previous metrics is to complete them with semantic features, thanks to word alignment techniques complemented using external knowledge bases like Wordnet (Hendrickx et al., 2009). Text entailment relies heavily on syntactic information that indicates if a text segment is implied by another one. The approach remains consistent by also using lexical alignment module but this time based on syntactic trees. On their side, Radev (2000) have first considered cross-sentence subsumption to check if one sentence implies another, to decide if it should be included. Finally, others the approach introduced in (Lloret et al., 2008) relies on a pretrained textual entailment classifier to measure sentences implication in summary candidates. Once the similarity is specified between all textual units, the selection of segments is usually done through a predefined threshold, but we also see that some authors favor more evolved techniques in order to fuse similar segments and thereby provide a better structure and approximation of true summaries (Barzilay and McKeown, 2005; Barzilay and Elhadad, 1999; Bing et al., 2015). Interestingly, for recent abstractive approach, pairwise similarity is used as a preprocessing step to better guarantee the non-redundancy of the phrases explored (Ghadimi and Beigy, 2022).

The methods we have introduced so far to evaluate redundancy consider the task of ranking the document as separate from estimating independence and similarity. This idea was first introduced by Carbonell and Goldstein (1998), who defined the task of maximal marginal relevance (MMR), in which each segment's score is directly penalized by its similarity with previously selected segments. However this definition of MMR is oriented toward query-based summarization, and thus some authors have designed methods using a feature-based importance score as a criterion to complement the redundancy penalty (Mori et al., 2005; Liu et al., 2006), or to adapt the model to multi-document summarization (Boudin et al., 2008). Finally, some authors have taken advantage of the richness of the representation provided by word and sentence embeddings to upgrade the similarity calculation employed in the MMR (Lamsiyah et al., 2021a; Chowdhury et al., 2022). Besides demonstrating better performance than strict similarity estimation (Xie and Liu, 2008), another benefit of this technique is that it creates summaries that are closely related to nonprofessional human summaries (Ribeiro and de Matos, 2007). Other methods have considered improving these rankings method by increasing the independence conditions between selected sentences in the ranking task through the shrinkage of text-segment representations (Yao et al., 2015), with methods such as pivoted-QR (Conroy et al., 2006), or by project sentences on distant boundaries of a similarity graph (Dong et al., 2020) thereby further reducing redundancy in the results. Some researchers have noticed that the MMR approach is an NP-hard task, and thus that using a greedy algorithm to solve it could still lead to sub-optimal solutions. Finally, modifications have been proposed making it a linear problem to solve it optimally with integer linear programming (McDonald, 2007; Gillick and Favre, 2009). Some authors also proposed to enrich the redundancy estimation with various semantic features to perform better at the task of update summarization (Mnasri et al., 2017). Similarly to this modification, other authors seeking optimal solutions have suggested submodular monotone functions (Fang et al., 2015) guaranteeing optimality with greedy algorithms, determinantal point process algorithms (Ghadimi and Beigy, 2022), or nonmonotone graph-based functions with a high probability of optimality (Lin et al., 2009; Lin and Bilmes, 2010). The same principle was then used for the MMR approach and has been proved efficient for identifying true and relevant information in the context of summarization for improving fake news detection (Kim and Ko, 2021). It has also been employed to manage the salience and the consistency of updated content for real-time streams of multiple tweets (Li et al., 2021).

Approaches calculating pairwise similarities, as presented before, are a first attempt to attain this objective, but the main problem is that if information is important, its score will compensate for any redundancy penalty imposed thereby allowing redundant sentences in the final output (Zhang et al., 2005). Consequently, the objective is to propose a method that automatically balances for importance estimation by penalizing it through a multiplication of both similarity and importance itself (Yin and Pei, 2015). In early experiments, researchers implemented a ranking algorithm that multiplied the probabilities of the relevance and novelty of terms to score and subsequently perform update summarization (Allan et al., 2001). This approach has also demonstrated its ability to increase the coverage of various dimensions of an event in news summarization task (McCreadie et al., 2018). Other authors attempting to avoid redundancy in this way have considered squaring the probabilities of already included terms to penalize their inclusion in the summary too many times (Nenkova and Vanderwende, 2005; Vanderwende et al., 2007). Other authors proposed to combine the estimation of relevance and novelty in an affinity graph-based context. The affinity ranking score relying on sentence connectivity is directly penalized by the similarity of selected nodes multiplied by their relevance (Yin and Pei, 2015; Zhang et al., 2005). The objective is to penalize nodes associated to the most important ones. Other authors examine the affinity graph methods either by creating sinkholes in the random walk process to disadvantage visiting closely related neighbors (Zhu et al., 2007); or by reinforcing previously explored ones, thereby decreasing the probability that the random walk ends on their neighbors (Mei et al., 2010). Finally, another strategy used the structure of the networks to encourage random walks to visit external nodes distant from the ones already selected (Amancio et al., 2012).

Coverage: Novelty as topic diversity

The previously introduced methods to enhance the information novelty in summaries rely on low redundancy, but to produce good outputs, they still need a better variety of information (Mei et al., 2010). Indeed, by guaranteeing independence between selected segments, these approaches do not ensure access to diverse content (Zhang et al., 2005). This notion of increasing novelty through diversity is especially justified in information retrieval theory because users prefer high-recall research results that will tend to support an extensive coverage of different topics (Zhai et al., 2015). Thus, the ideal summary can be considered as one presenting diverse information that covers as many aspects as possible of the original document(s). We can distinguish two distinct approaches favoring diversity in the outcome: one that implicitly models diversity in the content ranking process and one that explicitly tries to maximize the variety of content coverage in the summary.

The first manner to diversify information is to force the method to include all the multiple topics addressed in the original document(s). The supposition is that each cluster of related text segments will deal with the same aspects of the original document(s) (Hatzivassiloglou et al., 2001) thus having a high probability of similar and overlapping content (Abu-Jbara and Radev, 2011). Once clusters are created, it is the sentence-selection method that is interesting and makes it possible to avoid redundancy. The objective is thus to create topic clusters then design strategies to minimize the number of sentences picked for each cluster (Chowdhury et al., 2021). The first methods select top-ranked segments of each cluster (Banerjee et al., 2016; Abu-Jbara and Radev, 2011; Harabagiu et al., 2007) and ensure that the different relevant clusters are explored and that the number of clusters is determined by the number of sentences that should be included in the summary (Radev et al., 2004b). Other authors use complementary similarity measures such as cosine or Normalized Google distances (Alguliev et al., 2009; Song et al., 2011) to guarantee the quality of the ranked sentences per cluster. Clusters of topically related documents can also be used as an input for summarization models to make sure to induce diverse outputs. Chowdhury et al. (2022) employs this approach upstream to a graph-based model to assess relevance. For abstractive approach, the same technique is applied to provide various clusters to generative algorithms to guarantee the coverage of diverse segments. In the context of opinion summarization Pecar (2018) first propose to cluster opinionated features based on aspect detection to ensure the coverage of different products aspect. Coavoux et al. (2019) and Bražinskas et al. (2019) also create clusters and used a trained language model to establish distinct consensual abstractive summaries for each aspect of customer reviews and thus diversify their output. Another related technique consists in fusing similar sentences into clusters to only select the most relevant ones as leave-one-out strategy to create pseudo-summaries for fine-tuning models (Suhara et al., 2020). Angelidis et al. (2021) dynamically learn aspect cluster through multi-head latent representation during the training of a language model, then they use them to orient the summary generation towards diverse aspects in the summary. Finally, topic models have also been used to promote textual diversity in summaries. Gong and Liu (2001) were the first authors to exploit this idea in their work. They create their topic representation through the SVD decomposition, they add a constraint in the selection process to include a different topic each time. Some authors pursued this concept of maximizing topic coverage with submodular optimization (Shang et al., 2018) or bipartite word/topic graph (Parveen et al., 2015) to explore topics as much as possible.

One issue with these solutions is that they only consider global diversity, which does not guarantee the expansion of thematic content at the sub-document level. One way to encourage diversity directly and explicitly is then to maximize coverage of the concept and the semantic volume of the original document(s) (Yogatama et al., 2015). The authors have chosen to create an intermediate representation of the segments with deep learning models. By maximizing the distance between sentence embeddings, this tends to favor diversity through semantic volume (Yogatama et al., 2015). Several researchers have followed and modified this idea when applying deep learning techniques (Cheung and Penn, 2014; Cao et al., 2015). Finally, this approach was adapted for the LDA method by modeling two distributions, one for the previously seen information and one for the new. We then maximize the difference in selected elements in the context of update summarization (Delort and Alfonseca, 2012). The approach deployed by Filatova and Hatzivassiloglou (2004) tries to cover as many conceptual units as possible by formulating the problem as a maximum coverage knapsack constraint solved by a greedy algorithm. Diversity is ensured by rewarding the number of different units while setting a constraint to penalize sentences not containing enough novel units. This greedy method has been improved over time especially to assure getting an optimal solution (Takamura and Okumura, 2009). Instead of considering conceptual units, we can employ bigrams and impose the constraint directly to have elements included in the summary only once (Gillick et al., 2008). Liang et al. (2021), for their part, propose maximizing the diversity of elements by promoting sentences that maximize the coverage of all aspects of the input documents in a graph model. Finally, the authors in (Lin and Bilmes, 2011) have been able to create a monotonous submodular function including diversity constraint that has a constant factor to guarantee optimality. In the context of abstractive summarization, two strategies have been adopted to maximize the expansion of the semantic space. The first technique conditions text generation with seeded aspects to train a language model, then authors or end-users can devise strategies to enforce the inclusion diverse aspects by the text generator to explore different contexts (Amplayo et al., 2021). Otherwise, the approach consists in randomly leaving out different aspects sentences from input as a fine-tuning strategy of a large language model, allowing it to focus on different facets of the text when generating the summary (Oved and Levy, 2021).

Conclusion

While both redundancy and diversity attempt to increase the coverage of the input included in the summary to improve its quality, these two notions diverge in the type of information they provide. Indeed, with information redundancy, we can estimate the local importance of content within a collection of similar text segments, whereas with diversity, we seek to evaluate this quantity at the global level of the corpus and between different topics (Dong et al., 2020). By minimizing the information repetition, we are trying to gain information, but we do not guarantee to cover a whole set of topics (Zhang et al., 2005). Conversely, by increasing diversity, we ensure that information from all topics is embedded, however, as the same information can be linked to two distinct topics, we don't necessarily guarantee a gain in new information. This is why methods enforcing diversity, such as extracting one sentence from different clusters or several topics, are not efficient for update summarization tasks where information is concentrated around the evolution of a specific event (Zopf et al., 2016; Delort and Alfonseca, 2012). Whereas the same diversity-maximizing approaches perform extremely well when the objective is to cover various events in a news stream (McCreadie et al., 2018). Once again we see emerging the intimate link between novelty relevance, summary generation steps, and purpose factors and how considering understand fundamental characteristics of certain approaches.

Assessing unsupervised summarization methods

In order to better understand the influence of topical and novelty relevance, it is essential to contextualize the use of unsupervised document summarization methods. Presenting the resources and data sets available and the used evaluation metrics thus makes it possible to describe how the community takes relevance into account, and how it affects the evolution of approaches.

Resources and data sets

Increased involvement in automatic text summarization due to the proliferation of available data on the internet thus drives the interest in having concrete common and standardized resources to analyze the different approaches in real contexts. Since the evaluation of natural language processing methods, as in most fields related to machine learning, is done through the comparison of systems with each other, that input data is of a major influence in the production of summaries (Hovy et al., 1999; Jones et al., 1999), and that many solutions are specialized for certain types of data. It is therefore essential to obtain a global view of the characteristics of the data sets used to understand the fundamental differences explaining the various behaviors of unsupervised systems. Note that the objective of this section is not to provide an exhaustive review of the data sets used in the summarization literature as several good reviews already exist on this topic (Mani, 2001; Dernoncourt et al., 2018). The objective is to rather present data sets that are either used in unsupervised summarization and to provide enough material to allow a discussion on the potential biases it can generate in the design of unsupervised systems.

The first conference dedicated to automatic text summarization, that took place the first time in 1998, was the TIPSTER Text Summarization Evaluation SUMMAC (`https://www-nlpir.nist.gov/related_projects/tipster_summac/`). In addition to the evaluation framework, which will be discussed in the next section, the organizers made a data set available taken from newspaper sources. The data set is made of 20 topic-related collections each containing 50 documents selected from the top 2000 results returned by queries from an information retrieval system. The task provided was to constitute two summaries, one of a fixed length of maximum 10% of the original document size, and the other with no size constraints. Another conference, The National Institute for Informatics Test Collection for IR3 (http://research.nii.ac.jp/ntcir/outline/prop-en.html), also provided a news-based data set in Japanese. The objective was the production of both extractive and abstractive summaries of single news articles. Beside these two conferences, another major conference for text summarization was the Document Understanding Conference, DUC (https://duc.nist.gov/), which took place yearly from 2001 to 2007. During the first years of these conferences, the objective was to produce generic summaries of single and multiple documents. For both challenges the data set consisted of 30 document sets of 10 news stories each, for which three human annotators constructed different summaries of 50, 100, 200, and 400 words. Since 2005, the data set has evolved towards user-oriented applications. The tasks were once again based on news-story sets but focused on topic, query, viewpoint, or event, to facilitate comprehension of the expected assignment and so participants could concentrate their efforts in the same direction. In its final year, DUC proposed another evolution by creating a summarization updating task consisting of creating an output, knowing that the user has already seen documents answering its information needs. Once again, for each of these task, human annotators provided different summaries of up to 100 words. As of 2007, DUC conferences are no longer organized and have been integrated into the Text Analysis Conferences (TAC) (https://tac.nist.gov/about/index.html). Since 2008, the TAC has pursued the diversification of summarization projects. In 2008, they continued the news story summarization updates, but also added new data sets based on opinion blogs. In 2009 and 2010, TAC reoriented the analyses toward news stories but increased the diversity of the various challenges by adding one task dedicated to the evaluation of summarization systems and another to guided summarization, where the user's need is predefined to guide the method. For further details about the data, the tasks, and the objective of these conferences we refer the readers to the following papers (Mani et al., 1999; Over et al., 2007; Dang and Owczarzak, 2008). These conferences have been very useful to provide a normalized control framework for the community, and the data sets continue to be improved, enriched, and employed by current researchers to analyze the performance of their systems. As we can see, the tasks were particularly oriented to summarizing newspaper-based data. Because of the easy access of this information on the internet, several other data sets have been used in unsupervised automatic summarization over the years. The first data set we can describe is the Reuters news corpus (`http://boardwatch.internet.com/mag/95/oct/bwm9.html`), composed of 1000 documents and their associated extracted sentences, which represent approximately 20% of the original document size. Another frequently used data set is the TeMario Corpus (Pardo and Rino, 2003), which is constituted of 100 Portuguese documents (60,000 words total) extracted from Brazilian newspapers on several topics along with their good quality abstract summaries. The final notable sources of news stories used for automatic text summarization are the CNN & DailyMail corpus (Hermann et al., 2015), made up of roughly a million news stories and human-written abstractive summaries, which are related to a specific query; the Multi-News corpus (Fabbri et al., 2019) made for multi-document summarization and composed of more than 250 000 paired news and summaries; and finally the NewsRoom corpus (Grusky et al., 2018) composed of 1.3 Million extractive paired summaries that aims to measure inclusion of diversity and novelty by automatic systems. It is also worth noting that many authors have designed their own data set for their experiments that fulfill the guidelines for control, provided by the different conferences (Zhang et al., 2005; Clarke and Lapata, 2008). Following the TAC 2008 recommendations, some authors look to other source documents to evaluate their systems in order to diversify the systems created and their properties. The first historical data sets used were scientific papers collected for the purpose of single document summarization. Luhn (1958) was the first to attempt to generate abstract of papers, and he has been followed by several others, where either extracted sentences or an abstract texts are provided and annotated by human judges for various fields such as chemistry (Edmundson, 1969), computer science (Barzilay and Elhadad, 1999), or medicine (Parveen et al., 2015). Another well-studied set of data is the opinion data extracted from blog sources directly following TAC 2008 or from customer reviews available on the web. The most famous data set for this purpose is the Opinosis data set (Ganesan et al., 2010), which is composed of 50 topic reviews of hotels, cars, and various products and which takes redundant reviews related to queries. Each topic contains 50 to 575 sentences and 1- to 3-sentence summaries produced by human experts. Another opinion corpus is the Yelp Dataset Challenge and Amazon reviews (McAuley et al., 2015), which is specialized in abstract summarization of product reviews. Several authors have then proposed annotated corpus for summary evaluation based on set of 8 reviews per products (Chu and Liu, 2019; Bražinskas et al., 2019). OPOSUM is another dataset based on Amazon products (Angelidis and Lapata, 2018). The dataset aims to provided human selected references that provides important aspects and extractive summaries focusing on slaience, popularity, fluency, and redundancy. (Angelidis et al., 2021) have recently introduced SPACE, a corpus composed of 1.1 Million reviews based on TripAdvisor hotel reviews with manually hand-crafted abstractive summaries. The pupose of this dataset is to focus on aspect-specific summarization. Other authors have used diverse opinion data sets such as the TAC blog data set (Ferreira et al., 2013); IMDB movie reviews , which was in fact originally created for sentiment analysis (Denil et al., 2014); or even manually designed ones. The last common kind of data used for automatic text summarization is meeting transcripts. The ISCI corpus (Janin et al., 2003) consists of 75 transcripts of naturally occurring meetings, where human annotators were asked to write 200-word abstractive and extractive summaries. The AMI Corpus (Carletta et al., 2005) consists of 19 scenario-based meetings in which participants were asked to design a new product. These meetings have also been transcribed and annotated by human experts, once again in 200-word abstractive and extractive summaries. Finally, we can cite other types of sources that have been used only marginally, such as single summarization of books (Ceylan and Mihalcea, 2009; Cristea et al., 2005), emails (Yousefi-Azar and Hamey, 2017), banking reports (Dohare et al., 2017), or Wikipedia articles (Alami et al., 2018; Al-Radaideh and Bataineh, 2018).

For the 137 papers included in this review, Table 0.1 presents a distribution of the different corpora used in the evaluation of approaches. The different categories analyzed in these papers are news-based documents, opinion or blog data, scientific papers, meeting

Table 0.1: Distribution of the data sets

Data sets	Number of papers
News articles	112
Opinion data and blogs	19
Scientific papers	13
Meeting corpora	9
Others	17

corpora, and others. If systems were applied on multiple corpora for evaluation, we count one occurrence for each category.

We can see that despite the recommendations previously made by the TAC conferences to increase the variety of data sets used in order to increase the diversity of systems and possible applications, more than 70% of evaluations are still performed on news information. News sources are highly formatted, using a inverse pyramid structure where most important information is presented at the beginning of the document, and are well written, often focusing on specific events and providing answers to specific questions: *who, what, where, when, why* (Owczarzak and Dang, 2011). These authors also note that there is broad consensus on the reported facts through multiple documents, creating a homogeneous distribution of terms. The very specific attributes of this type of data might create limitations in the design of approaches, especially in unsupervised summarization, where the algorithms are very sensitive to the implicit characteristics of the data. These limitations are reinforced by evaluation measures, as will be discussed in the following section, but we understand the need to provide a common environment, especially in the beginning of the discipline, in order to facilitate participants' understanding of the expected objectives and the analysis of the strengths and weaknesses of their algorithms through controlled evaluation procedures.

Evaluation approaches

While the impact of information novelty and topical is important to understand the major differences in the functioning of a model, it is essential to compare them on a similar basis to understand the effect of those purpose mechanisms and how they impact performances. One of the best ways to get feedback on method designs is by evaluating of outcomes. However numerous measures propose several properties for analyzing summary content and can differ greatly or even appear inconsistent in their definition and interpretation. These metrics will therefore also present flaws and biases toward information and, especially how it is encapsulated (Bhandari et al., 2020; Fabbri et al., 2021). For example, as mentioned before, the influence of datasets, but also the diversity of acceptable solutions, makes it more difficult to determine what material should be included in the final summary (Narayan et al., 2019). Therefore, the objective of this section is not to provide an exhaustive review of the performance metrics used in summarization literature as several good works already exist on this topic (Deutsch and Roth, 2020; Fabbri et al., 2021; Wu et al., 2020). It is rather about demonstrating what standards are applied to the analysis and evaluation of unsupervised methods nowadays. The final objective of introducing those metrics is to justify the choices in evaluation made in this thesis, but also to discuss how they are related to topical and novelty relevance, the purpose factors, and how it influences the perception and the design of summarizers.

Intrinsic evaluation

Among the methods for evaluating intrinsic performance, there is a distinction between two categories for assessing the quality of a summary (Mani, 2001). The first concerns its informativeness, where we measure the fidelity of the output to the original documents, and the second is the quality of the production, where we judge the coherence and how well the summary can be read.

The target of automatic text summarization is human users, thus it is very natural to compare such systems to human productions. Thus, in order to evaluate the fidelity of the final summary to its source, most of the current evaluation methods use reference texts created by people as the gold standard. Even if there is a major difficulty with this approach, due to the fact that two very different summaries can be considered as good in terms of their summarization, once the purpose of the task is clearly defined and multiple humans are asked to produce a reference, the space of possible output is greatly reduced (Mani, 2001). Thus it become acceptable to use these gold standards to judge automatic systems and compare their performances. The first used evaluation technique was performed manually by people: Selected experts were asked to judge , on scales of one to five, whether the included text segments convey the information contained in the source. One famous framework for this type of evaluation is the Summary Evaluation Environment (SEE) (Lin, 2001) used in the first years of the DUC conferences (Dang, 2006). The framework provides an interface to compare a reference document to the peer summaries and annotate the pertinence of each text segment. However this method suffers from the variability of human judgments and it is extremely time consuming. With access to these gold standard documents, informativeness is quite easy to assess through automatic processes. The first automatic evaluation metric proposed for automatic text summarization is an adaptation the well-known metric of the information retrieval task: precision, recall, and F-measure (Lloret and Palomar, 2012). This metric is especially useful for extractive summarization, where precision represents the proportion of sentences correctly selected by the system, recall is the proportion of sentences selected by judges and selected by the systems, and the F-measure is the mean of the previous two. However, this method still suffers greatly from the variability of the created gold standards. The work of Radev and Tam (2003) demonstrates that humans tend to agree more when it comes to ranking important segments to include in the summary. Thus, they propose the Relative Utility score, where we compare the rankings provided by experts to the ones predicted by the system. While not agreeing deeply on whole sentences to integrate in summaries, human evaluators still agree on most the important terms to include (Nenkova and Passonneau, 2004). The Pyramid method exploits this property by considering Summary Content Units (SCUs), which are pieces of information that overlap in different human summaries, as worthy to include in the final output. Then it measures the proportion of SCUs contained in each system to assess its quality. This metric makes it possible to obtain valuable information on the analyzed approaches and has been adopted in the DUC and TAC conferences. But, the annotation of the SCUs still requires a huge amount of manual time and effort. In order to propose a fully automatic evaluation procedure, Lin (2004) introduce the Recall Oriented Understudy for Gisting Evaluation (ROUGE) score, which is an adaptation of the BLEU score used in machine translation. This score is an approximation of the recall measure but is based on the proportion of *n*-grams overlapping between the gold standard and the automatic summary. Several variations of this metric exist, including the classic ROUGE-N, which directly measures the overlap of *n*-grams, the ROUGE-L, which measures the longest common subsequence, thus taking into account word order; and the ROUGE-W, which weights the sequences with the number of direct consecutive words. These metrics have been extremely used in conferences and papers, especially because the author shows that it correlates well with human judgment. One weakness of ROUGE is that it only considers strict *n*-gram matches, thus some authors propose the Basic Elements (BE) metric (Hovy et al., 2006), which instead considers the proportion of relation triplets (head|modifier|relation) between references and system outputs. This metric demonstrates greater flexibility in evaluation because it allows matching equivalent expressions that do not contain the exact same words. Once again this metric has been extensively used for evaluation in the DUC and TAC conferences. However, due to the variety of existing potential solutions to form the gold reference, selecting one solution as the valid summary presents a reference bias (Louis and Nenkova, 2013). A recent approach thus proposes to overcome this bias by annotating, via multiple non-expert judges, the relevant content directly in the input document(s) (Narayan et al., 2019). Once this new labeling is done, we can then use our traditional evaluation methods such as precision, recall or ROUGE to obtain scores that do not penalize summaries containing information different from the single reference chosen.

Extrinsic evaluation

The rise in the amount of textual data available obviously creates issues for humans to digest information but it also leads trouble for other systems because of the increased amounts of noise and time needed to compute this quantity of material. Automatic summarization is seen as a way to solve these issues. Extrinsic evaluation processes aspire to assess the efficiency of these automatic productions. These metrics offer different advantages over intrinsic evaluation because the variety of tasks and objectives that can be used to judge summaries increases the richness of the analyses of such systems, and because these tasks are related to real industrial applications and to the information needs of end users (Mani et al., 1999).

The first criterion to assess the usefulness of a given summary is to observe whether it fulfills specific user information requirements. One way to determine if the summary can respond to some information need is to see if the final output provides sufficient material to relate it to the same topic as the original document. This specific task is defined as relevance assessment (Mani, 2001), where the accuracy and execution time of an ad hoc system are evaluated with initial documents and a summary. The first task submitted for this type of evaluation by is the categorization game (Hovy et al., 1999), where the methods infer a topic category for the original document and the summary, and the evaluation consists in measuring the correspondence between both classifications. Another task used early on for evaluation in conferences was question answering, not to be confused with the question game presented below, because it models concrete activities. The objective is to ask a question as an input of the system and observe whether the output produced includes elements of the initial documents that are considered parts of the answer if any. The recent work with *APES* (Eyal et al., 2019) and the work proposed in (Scialom et al., 2019) pursue this idea. The authors show that by implementing an external pre-trained question answering system based on deep learning techniques, they obtain a metric that displays good correlation with the Pyramid score (Nenkova and Passonneau, 2004) and human evaluators without necessitating labeled data. Finally, another way to assess the relevance of the document is through information retrieval tasks, where the purpose is to measure if recovered summaries are ranked the same way as the original inputs (Nomoto and Matsumoto, 2001b) or if the returned results correspond to the topic defined in the input query made by the user. Another kind of extrinsic task that can be designed to assess document quality directly relates to the notions of the informativeness and fidelity of the source documents, as previously discussed. These are reading comprehension tasks (Mani, 2001) where the goal is to evaluate how much information from the original document is conveyed by the summary. The first introduced tasks relative to these categories of metrics are the Shannon Game and the Question Game (Hovy et al., 1999). The Shannon task aims to impute an information content score to terms from the document and the summary in terms of how they make it possible to determine the overall message. Then, if a summary includes most of these terms, it is easy for a human to reconstruct the original input by reading only the output because the elements are informative. The Question Game consists of asking multiple-choice questions to users about the document content. Then the correctness of the answers is measured via different frameworks: if the readers have seen the initial corpus, if they have only read the summaries, or if they have viewed both. It allows to understand how well the summary replaces the most important facts conveyed by the input and how suited it is as an alternative source of information. These tasks essentially measure the extent to which the information in the original documents has been covered. Some authors have then decided to introduce metrics to assess this coverage. First authors propose heuristics such as measuring the Jensen-Shannon divergence directly between the summary and the original documents (Louis and Nenkova, 2013). However, these results do not show a good correlation with human productions. Other approaches such as *BertScore* (Zhang et al., 2019) or *SUPERT* (Gao et al., 2020) have improved these results by measuring the *Word Mover's Distance* (Kusner et al., 2015) between embedding representations of *n*-grams and by using alignment techniques between the summary and the original documents (or extracts of them). These new extrinsic measures properly outline the extent to which there is an overlap of the information contained between the source and the summary, thus coming closer to the definition of the information coverage relevance measures as first specified in (Mani, 2001).

Comparative analysis of evaluation metrics

As we can see, there are many different methods that compare and analyze the different summarization techniques proposed in the literature (Lloret and Palomar, 2012; Rankel

Table 0.2: Distribution of the evaluation metrics

Metric	Number of papers
Human Evaluation	32
Quality and Grammatical Properties	12
Relative Utility	3
Pyramid Score	10
Precision, Recall, and F-score	35
ROUGE	137
Basic Elements	2
Classification Game	4
Question Answering Tasks	2
Information Retrieval Tasks	2
Shannon Game	1
Question Game	0

et al., 2013) since they have not been applied in the analyzed papers of this review. The evaluation process is a very difficult task where no consensus has been found, because each method has its own strengths and weaknesses; thus this multiplicity presents strong opportunities for the community. However, it is interesting to note that historically there are no metrics dedicated to unsupervised automatic summarization even if the trend seems to improve with the emergence of new extrinsic metrics (Fabbri et al., 2021). Once again, we perform a quantitative analysis of the different metrics used in the papers covered in this literature. The results of this analysis are displayed in Table 0.2 below.

These results clearly demonstrate that most of the methods employed in the literature for evaluating systems are intrinsic approaches, and that most of these rely on the production of gold-standard documents. Even in the intrinsic methods, we can clearly see that two techniques account for most of the evaluation methods. Another aspect that should be noted is that when the ROUGE score is used for evaluation, approximately 70% of the time the metric is used alone, and when it is employed with other metrics, it is mostly used with other automatic intrinsic evaluation methods such as pyramid score or precision and recall (60% of the time), and is only used 45% of the time with complementary quality metrics evaluated by human experts. This tendency is even stronger when we analyze its application through time, since the ROUGE score has increasingly been used in recent studies. Our results also correlate with the same tendencies observe on the use of only one metrics correlating human evaluators in recent papers by Steen and Markert (2021); Fabbri et al. (2021); Narayan et al. (2019). However, it is essential to note that most of the recent papers followed the recommendations made in these reviews and provide a complementary analysis by human reviewers on various dimensions such as salience, consistency, factual coherence, in addition to more traditional factors on grammatical fluency. This being said, evaluations concerning salience or consistency still pose issues, since the definition of the latter is rarely provided to understand which notion is being evaluated. All the more so as it is now well known that assessing the relevance of information is extremely complex for human evaluators when the latter are not sufficiently constrained (Kryściński et al., 2019). Once again, we agree that this homogenization has enabled many advances in the early days of automatic document summarization because it provides a clear basis for the comparison of systems and definite parameters for analyzing automatic summarization systems. However, it is now widely accepted that there are two definitions of abstract quality: coverage and informativeness (Narayan et al., 2019). The perception of this quality is also directly influenced by the fundamental notions of information relevance, either topical or novel. We have examined the influence of these notions on the creation of unsupervised systems. The observation of the current trends on the evaluation of these methods will allow us to bring a discussion on the long-term impact of these choices can have on the unsupervised approaches of automatic document summarization.

0.1 Conclusion of the theoretical framework

A document summarization system must go through three main steps: Representing the input documents, scoring their content, and finally selecting it to generate a reduced-size text that meets a specific task and an information need. Each summary is made up of several dimensions that characterize it in terms of input, output and the purpose it is fulfilling. This purpose factor then distinguishes the summary usage and the audience to which it will be addressed. These different dimensions will then modify the way information is encoded in the three main steps of the model design. The notions of selective information for salience or representative information for centrality, or the concepts of non-redundancy for information gain, or information diversity for coverage are all different ways of encapsulating relevance and thus meeting these needs and tasks differently. Indeed, the producing a short summary for newspaper readers requires conciseness and precise information on the main topic or event. The ideas of a specific audience and an indicative usage is emerging as a purpose for the expected output. Favoring selectivity in representation and scoring, and non-redundancy in scoring and generation, appears to be the most appropriate manner of fulfilling this requirement. On the contrary, if the aim is to offer users a summary of opinions that will let them avoid reading all the other opinions, then an informative and general summary covering as many topics as possible should be proposed. Representing and scoring information to bring out centrality and generating text to privilege diversity seem the best ways to meet this need this time.

To confirm these intuitions, performance evaluation datasets and metrics become essential to compare and perceive the advantages and shortcomings of these systems. However, the evaluation of these models is extremely dependent on the task, the dataset, and on the metric itself, as they are not all intended to consider the same dimensions. Applying reference standards such as ROUGE provides an initial basis for comparison and understanding for known dimensions. But, knowing the bias of these metrics, it becomes crucial to implement metrics related to the defined usage, task, and audience. It is also indispensable to explicit the conditions and the intentions behind the dataset creation to understand its characteristics (Over et al., 2007). Therefore, it is essential to specify purpose factors, since this is what implicitly influences the whole design of the model, to fairly compete to other approaches.

In the remainder of this thesis, we will present three cases that evidence the importance of the audience and usage dimensions. Indeed, for each of the papers, we will modify a generalist model to adapt it to a task and user need we are trying to address. We also propose evaluation methods adapted to these defined needs. This allows us to demonstrate that our approaches are more efficient and better meet these demands.

Chapter 1

Unsupervised update summarization of news events

Published Article: Carichon, F., Fettu, F., & Caporossi, G. (2023). Unsupervised update summarization of news events. Pattern Recognition, 144, 109839.

Abstract

A long-running event represents a continuous stream of information on a given topic, such as natural disasters, stock market updates, or even ongoing customer relationship. These news stories include hundreds of individual, time-dependent texts. Simultaneously, new technologies have profoundly transformed the way we consume information. The need to obtain quick, relevant, and digest updates continuously has become a crucial issue and creates new challenges for the task of automatic document summarization. To that end, we introduce an innovative unsupervised method based on two competing sequence-to-sequence models to produce short updated summaries. The proposed architecture relies on several parameters to balance the outputs from the two autoencoders. This relation enables the overall model to correlate generated summaries with relevant information coming from both current and previous news iterations. Depending on the model configuration,

we are then able to control the novelty or the consistency of terms included in generated summaries. We evaluate our method on a modified version of the TREC 2013, 2014, and 2015 datasets to track continuous events from a single source. We not only achieve state-of-the-art performance similar to other more complex unsupervised sentence compression approaches, but also influence the information included in the model in the summaries.

1.1 Introduction

Automatic text summarization is the process of distilling information contained in one or more documents to produce a reduced version that meets the need of a particular task or user. The most frequently used data source in document summarization is news events. Because news events are inherently time-sensitive (Goldstein et al., 2000), update summarization was one of the first tasks emerging in the 2008 Document Understanding Conferences [1] (DUC) and subsequently taken up by the Text Analysis Conferences [2] (TAC) and the Text REtrieval Conference [3] (TREC) in the temporal and real time summarization tracks. Of course, news information is now consumed via new media. To wit, people are increasingly turning to blogs, web journals and Twitter for their news to, for example, keep up to date on developing events such as natural disasters (Rudra et al., 2018). Consumers, especially young people, are also increasingly likely to follow the news live through their smartphones. Specifically, more than 55% of smartphone users receive notifications on their phones alerting them to breaking news or major events. More crucially, half of these users will consult the full article after reading a notification [4]. Notification quality is then crucial because it provides some basic information to users who opt to not read the full article, while simultaneously increasing the likelihood that users will click though. In this regard, the notifications act as an efficient real-time summary of a given event that is further fleshed out with each subsequent short headline. The objective of update summarization

[1] `https://duc.nist.gov/`

[2] `https://tac.nist.gov/about/index.html`

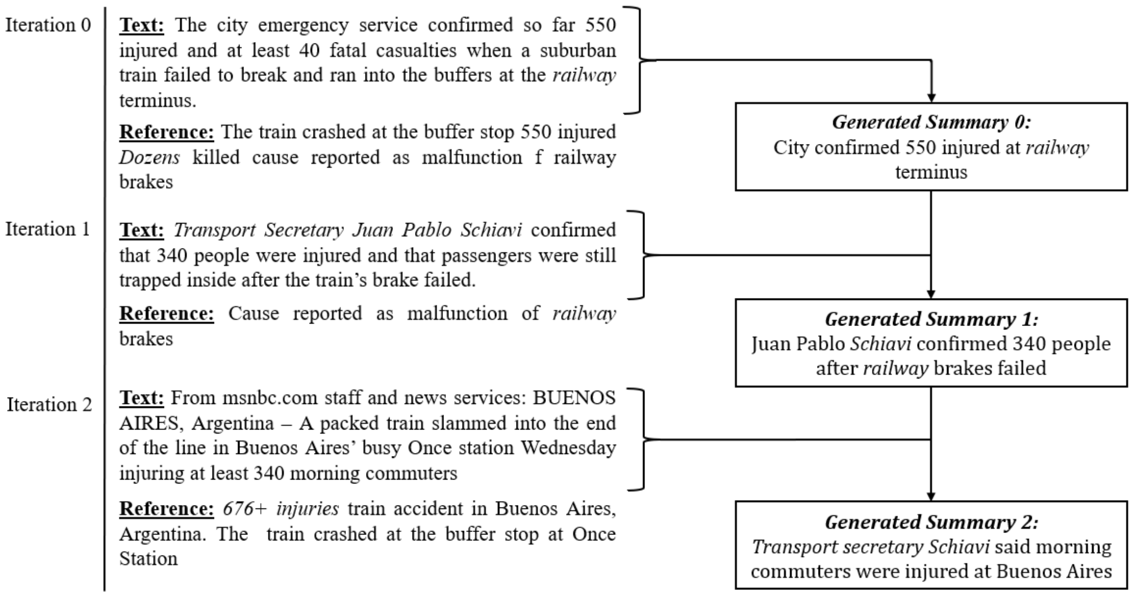
[3] `https://trec.nist.gov/`

[4] `http://pewrsr.ch/2ccvrnC`

is to produce outputs that include both relevant and new content that factors in some background knowledge, represented by previously generated material (Allan et al., 2001). For events lasting for longer periods, that background information is iteratively enriched with new data (Bysani, 2010; McCreadie et al., 2014). Thus, relevance and novelty are reassessed with every update to determine what information to include in the summary (Kedzie et al., 2015). The most efficient methods then consist in simultaneously scoring and selecting sentences depending on relevance and novelty to meet the user's need for information while guaranteeing the independence and quality of the sentences incorporated in the results (Agrawal et al., 2009; Carbonell and Goldstein, 1998). However, most of these methods ease the characterization salient and novel content through the analysis of redundant information in multiple sources.

In this paper, we introduce a new method to conduct update summarization for short single documents. Our approach has several advantages over existing methods. Since the documents and summaries are single short texts, we can't rely on classic techniques based on redundancy to identify salient content. Indeed, these approaches are mobilized at the news story level and thus require local redundancy of information within this news to estimate the relevance of the terms. They then define novelty as any material that is not redundant (Kedzie et al., 2015; Delort and Alfonseca, 2012). We also depend on local information to measure relevance and novelty, but we combine and constrain it with global metrics such as the Term-Frequency/Inverse Document Frequency (TFIDF). Crucially, this reduces the need for local redundancy and therefore makes it possible to handle single short documents. Moreover, these methods are generally designed for multiple documents and are thus extractive, which prevents them from producing coherent outputs in this context. Neural network systems applied to tasks such as update or progressive update summarization have proven to efficiently create abstractive summaries of single documents (Li et al., 2016). These approaches rely on the model capacity to identify relevant and new information in supervised fashion through the huge quantity of labeled samples. However, the example in 1.1 illustrates that this data is acutely noise, which will prevents supervised models from functioning as intended. Furthermore, in many real-world scenarios, the diversity of topics and genres of the data compromises access to the labels required to train those systems. Therefore, we assume it is better to increase the portability of our model by favoring unsupervised techniques. Recent neural network models such as autoencoders have demonstrated their efficiency to capture important content and generate unsupervised abstractive summaries (Chu and Liu, 2019). The main matter remains to find means to constrain the content and the length of the produced summary. Unsupervised approaches have been adapted to process short inputs by combining information selection techniques with semi-abstractive text generation (Févry and Phang, 2018; Baziotis et al., 2019). Despite these very promising results, none of these methods account for past summaries for a given event. However, the reference examples employed in the figure 1.1 point up the importance of considering this temporality to ensure the summary relevance and cohesion for the whole news story. For this reason, and as we can see, our approach compares the new information with the content it has previously generated. Moreover, we decide to use the summary as an input instead of the original text in light of extant research exposing that only half of users consult the full article after receiving a notification.

This paper builds on the previous work done on the TREC temporal track of datasets. Once again, most approaches to this task generate summaries through the analysis of redundant information in multiple sources (McCreadie et al., 2014). Therefore, to perform our task of update summarization on single documents, we combine the 2013, 2014, and 2015 TREC temporal summarization tracks and modify them to follow sources individually. The left part of figure 1.1 provides a sample news story (a train crash in Argentina) that could be extracted in this manner. It is also a case in point as to why relevant and updated live notifications are crucial for developing news stories. The news comes from a single source, and information is clearly reused through each iteration. For instance, the term "railway" is recycled from the original information in the first update summary to ensure consistency. We can also notice the application of novel terms such as "Argentina's transportation secretary" assumes the reader viewed the previous summary about the "Transport Secretary Juan Pablo Schiavi." The text on the right of figure 1.1 supports our intuition about the importance of allowing the model to control the novelty and/or



Iteration 0

Text: The city emergency service confirmed so far 550 injured and at least 40 fatal casualties when a suburban train failed to break and ran into the buffers at the *railway* terminus.

Reference: The train crashed at the buffer stop 550 injured *Dozens* killed cause reported as malfunction f railway brakes

Generated Summary 0: City confirmed 550 injured at *railway* terminus

Iteration 1

Text: *Transport Secretary Juan Pablo Schiavi* confirmed that 340 people were injured and that passengers were still trapped inside after the train's brake failed.

Reference: Cause reported as malfunction of *railway* brakes

Generated Summary 1: Juan Pablo *Schiavi* confirmed 340 people after *railway* brakes failed

Iteration 2

Text: From msnbc.com staff and news services: BUENOS AIRES, Argentina – A packed train slammed into the end of the line in Buenos Aires' busy Once station Wednesday injuring at least 340 morning commuters

Reference: *676+ injuries* train accident in Buenos Aires, Argentina. The train crashed at the buffer stop at Once Station

Generated Summary 2: *Transport secretary Schiavi* said morning commuters were injured at Buenos Aires

Figure 1.1: Data sample illustration. The left side introduce a news event concerning a train crash in Argentina. The event is updated 3 times. The *italicized text* in the left-hand column highlights which terms can be reused through iterations. It also displays discrepancies between sources and reference summaries. The text on the right-hand column demonstrates how our model uses generated content to produce improved iterative summaries.

consistency of information through the iterations to include content that would not be available otherwise. In addition, one significant aspect that we can observe in the example is the discrepancy of data due to our modification for single update summarization. For instance, one reference text states the number of "676+ injuries"; information which cannot be retrieved from either the source text or the previous iteration. This phenomenon makes it difficult to recognize the origin of the content, and why the material has been included in the final summary.

In response to the above challenges, we present an unsupervised autoencoder model where the summary generation is influenced by the novelty and coherence of the information previously provided. More specifically, the model relies on a Sequence-to-Sequence (Seq2Seq) architecture that simultaneously learns a language model (LM) and follows an information constraint model (CM) composed by a reduced length version of text with only the best Term-Frequency/Inverse Document Frequency (TFIDF) scoring words. The LM receives an encoded Recurrent Neural Netowrk (RNN) representation of the concatenation of the previously generated summary and the current text. The objective function enforces learning to reconstruct the information from both sources. Then, the same RNN encoder produces a representation for the reduced version of the current text that contains its informative content. Terms are selected based on their TFIDF values and weighted by their occurrence in the previous summary. The objective function is to recreate this reduced text. This second term serves both to enforce the model's length constraint and to provide guidance to select information to include in the final summary. The inclusion in both parts of the prior content makes it possible for the model to decide what material from the previous steps will be retained. Finally, in the generation step, the model must choose between following the LM or the informative content.

As such, this paper makes the following contributions:

- We introduce the new task of update summarization of short single documents. This task can also be used to achieve update sentence compression if the document is composed of a single sentence.

- We propose a modification of the TREC temporal summarization dataset to perform and evaluate this task.

- We present a novel unsupervised semi-extractive summarization approach composed by two competing auto-encoding models to generate summaries of short documents. This combined structure selects the most likely terms from either a language model that reconstruct grammatical and coherent texts or a length constraining model that only enforces the selection of salient words.

- To perform update summarization, we have also introduced a new hyperparameter modifying the behavior of both models. First, the parameter is employed in the learning objective of the language model to reconstruct the current text while considering the previous generated summaries. In the information constraint model, we use this parameter to weight the score of the relevant terms of the current text by their occurrence in the precedent iterations.

We assess the performance of our model using the standard ROUGE (Lin, 2004) and SUPERT (Gao et al., 2020) document evaluation metrics. Based on these metrics, our model's performance is either equivalent or better than more complex unsupervised baselines as well as certain supervised baselines. We also conducted human evaluations to judge the relevance and the coherence of the summaries. The results show that our model encompasses more salient information and, despite being less grammatically coherent than human references, it nonetheless produces understandable and reasonable texts. Finally, using both manual and automatic methods, we estimates the novelty and consistency of the summaries produced by different approaches, demonstrating that our model can be used to control information included in the output.

1.2 Related work

This paper draws on three topics of automatic text summarization, described below:

1.2.1 Unsupervised summarization

Classical approaches of unsupervised summarization have prioritized extractive methods that optimize the selection of salient, representative, and diverse sentences form one or multiple texts. More recent approaches relying on sub-modular functions and hand-crafted features (Ghadimi and Beigy, 2020), or the use of pre-trained models and a determinantal point process technique to account for redundancy (Ghadimi and Beigy, 2022) have demonstrated the interest of creating precise optimization functions to meet these criteria in order to achieve strong task performance. The emergence of deep learning techniques led to the generalization of abstractive approaches, as these create entirely new texts as summaries. Unsupervised models need a limit to produce a reduced version of the original inputs. For multiple documents, the constraint is implicit. Indeed, the objective is to learn a representation of the set of documents which will be decoded as an average of the input (Chu and Liu, 2019). For single document summarization, the constraint must be explicit and applied to the size of the original or produced text. This method is also applicable for new semi-extractive approaches where the model learns to generate a compressed version of the input sentence. The objective is then to set a constraint such that the model drops non-informative words from the original sentence to form the summary. The first approach suggests using a denoising autoencoder where authors add additional grammatical non-informative words to the input and the model learns to omit them afterwards to create the compression (Févry and Phang, 2018). Since compression coerces the system to remove content, some methods modify the CM to help the model apprehend the salient information. The model is then forced to respect new constraints depending on the main topics (Baziotis et al., 2019) of the documents, improbable informative words (Malireddy et al., 2020), or the mutual information between the original sentence and the compression (West et al., 2019). In this article we also introduce a semi-extractive approach with constraints to facilitate information capture for the summary. It differs from previous work because we apply a pretrained TFIDF model, which shows very good performance for news stories. Moreover, this simple and more explicit model allows us to account for previously generated information by updating the original TFIDF score with the previous word occurrences.

1.2.2 Novelty and consistency

The novelty principle is a fundamental concept for many NLP tasks such as information retrieval, Q&A systems, recommender systems, or document summarization. The goal is to offer a sufficient variety of information to users so that at least one item meets their expectations (Agrawal et al., 2009). To date, the most influential method in the field remains the Maximum Margin Relevance (MMR) (Carbonell and Goldstein, 1998). It introduces a ranking algorithm that selects relevant sentences and penalizes them based on their similarity to the ones already included in the summary. The MMR approach operates at the sentence selection step, but novelty can be estimated through the three steps of the summarization process: scoring, selection, and summary generation (Bysani, 2010).

Recently, several approaches based on the Seq2Seq model have tried to implement novelty systems in the summarization process at the scoring level (Fabbri et al., 2019). The model that most closely resembles ours implements the Maximum Mutual Information (MMI) in the objective function of the model to generate diverse responses in a discussion (Li et al., 2016). The authors argue in favor of integrating the MMI in the model objective function since it can capture inter-sentence relations. Although novelty is crucial to ensure the relevance of information, the consistency of consecutive content is also of utmost importance to evaluate what to include in the summary. Multiple applications in video summarization have highlighted the fundamental role of similar contextual frames for summary generation (Zhao et al., 2018; Zhu et al., 2020). Another approach further demonstrated the importance of diversifying attention paid to the context to improve salience estimation (Li et al., 2021). Finally, a recent approach demonstrated that using unsupervised learning on temporal and contextual data combined with reinforce learning techniques is useful for iterative or dynamic outputs (Zhao et al., 2021). In the case of textual data, the consistency of information–characterized, for instance, by the presence of related lexical items and smooth semantic transitions–is also key to ensuring the users' comprehension of the summary (Barzilay and Elhadad, 1999). Recent methods have focused on guaranteeing sentence reordering in the summary based on information found in previous sentences to ensure these semantic associations (Mohammed and Al-Hameed, 2021). A summary emphasizing consistency will then highlight the aboutness and indicativity of the summary (Barzilay and Elhadad, 1999) while the novelty will promote the informativeness of the text (Goldstein et al., 2000). Our approach accounts for these potential relations by linking the current text and the previous model output. More specifically, we added the notion of redundancy in the scoring of informative words to preserve. The constraint input is composed only of the highest TFIDF scores, which is then weighted by the frequency obtained in the previous summary and by a MMR method for reconstructing the documents. The model is then able to select either the word in the language model that foster appropriate for new inter-sentence relations/associations or words that maintain consistent update information.

1.2.3 Update summarization

The goal of the update summarization task is to produce a summary that focuses on new relevant facts for users. The scope of the update can include single updates or the temporal tracking of a continuous stream of new incoming documents that are slated for summarization. It may also be a progressive creation of new material or the dynamic modification of the output at specific timed intervals (Bysani, 2010). As the system deals with an unfolding event, the information salience, which is often based on content redundancy, become too complex to evaluate over time (Kedzie et al., 2015). To facilitate this process, the event is considered as a set of documents, which opens the door to the application of classical multi-document summarization (MDS) methods. The major challenge is to come up with new material compared to previous iterations. In the first attempt, the authors implemented a ranking algorithm that multiplied the probabilities of the relevance and novelty of terms to score and subsequently include them in each update (Allan et al., 2001). The same principle was then used for the MMR approach and has been proved efficient for identifying true and relevant information in the context of summarization for improving fake news detection (Kim and Ko, 2021). Other techniques propose to dynamically weight the novelty and relevance thresholds in ranking systems using the quantity of information present in each period (McCreadie et al., 2014). Finally, instead of using similarity and redundancy, other authors have handled the task by diversifying the topics covered in the updated output by including information from different document clusters or topics (Delort and Alfonseca, 2012). However, all of these approaches overlap insofar as they introduce extractive multi-document summaries of the original content. Neural networks models such as Seq2Seq architectures make it possible to generate a new piece of text conditioned by temporal information. These abstractive models can create compressions applicable for both single and multiple documents summaries. Particularly, this method has been applied to single document summarization where the conditioning information is taken from an external source such as Wikipedia for knowledge transfer (Prabhumoye et al., 2019). It has also been employed to manage the salience and the consistency of updated content for real-time streams of multiple tweets (Lin et al., 2021). Our approach is also applied to single document update summarization; however, one key difference is that our approach is unsupervised and therefore more portable to different fields. Moreover, the use of frequency and TFIDF to estimate term importance make it possible to apply this model to short documents or to sentence compression where redundancy of information is scarce.

Our approach is the first to propose an unsupervised Seq2Seq model where the summary generation is conditioned on the information previously seen by the user. This method is not entirely abstractive; rather, it generates a compressed version of the same input sentence. Nevertheless, it allows us to demonstrate the applicability and portability of our model on new tasks such as the update of short text summarization where training data is scarce.

1.3 Proposed model

This section presents the general architecture of our suggested approach, as depicted in Figure 1.2. As seen in the figure, our model is composed of two competing autoencoders. The first autoencoder learns a language model (LM) to produce coherent texts, and the second autoencoder constrains information (CM) to force the model into create a shorter selection of relevant words to include in the summary.

In the section 1.3.1, we begin by introducing the classic architecture of the autoencoder used for both models. Then, in section 1.3.2, we present the design of our CM, and and place special emphasis on the formula that characterize word importance based on their TFIDF scores. It also details how that new input weights the score based on the frequency of previously generated terms. In section 1.3.3, we explain the modification made to the LM encoder so that it can consider information from previous iterations when reconstructing the current text. In section 1.3.4, we we go in greater depth on our personalized objective functions that mimic the behavior of maximum margin relevance approach as well as on how we account for the information constraint. Finally, in section 1.3.5, we detail how the two models compete against each other in order to produce a short, relevant and coherent summary at each iteration and how our new parameters influence novelty or consistency in the final outputs.

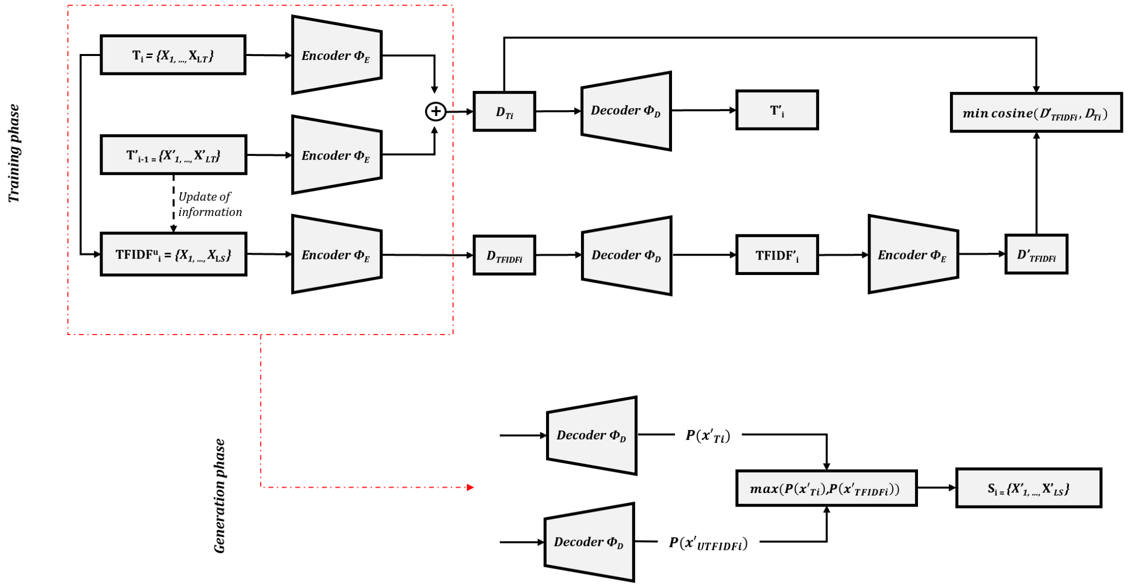


Figure 1.2: General model architecture. Autoencoder architecture for update summarization. At each text iteration T_i, we use as additional input the previously generated summary $T^{'}_{i-1}$ to produce the current representations D_{Ti} which is used to train the language model. We also use $T^{'}_{i-1}$ to generate $TFIDF^u_i$, the vector composed of the most relevant updated terms. This representation is encoded into D_{TFIDFi} to train the information constraint model. At the generation stage, we select the highest probability between $T^{'}_i$ or $TFIDF^{'}_i$ to form the final summary.

1.3.1 Model background

In this paper we use the classic encoder-decoder architecture for our documents. We introduce in this section the basic Seq2Seq model on which we rely. Let's note $T = \{T_1, ..., T_i, ..., T_c\}$ the corpus of c documents covering an event across time. Each document T_i corresponds to a specific iteration or update of the news story, and can be represented by a set of N_i words $X_i = \{x_1, x_2, ..., x_j, ..., x_{N_i}\}$. The model encoder produces, through the application of a bidirectional Gated Recurrent Unit (GRU) (Cho et al., 2014), a document encoding D_i and some encoder hidden states $h_1, h_2, ..., h_j, ..., h_{N_i}$. When decoding, we perform N_i decoding steps to generate our sentence. We start by fixing the initial hidden state of the decoder s_0 to the hidden representation of the document D_i, and, at each decoding step t, a simple GRU decoder estimates the current hidden state s_t with the states s_{t-1} and the predicted word $x^{'}_{t-1}$ at preceding steps:

$$s_t = GRU(s_{t-1}, x^{'}_{t-1}) \tag{1.1}$$

At decoding step t, the energy vector e^j_t of each input words j and the attention distribution a_t are named and calculated this way in this article as in (Bahdanau et al., 2014):

$$e^j_t = v^\top \tanh(W_h h_j, W_s s_t, b_{attn}) \tag{1.2}$$

where v, W_h, W_s, b_{attn} are learnable parameters of the model. We thus obtain an energy vector e_t over the whole input text for each decoding step. The attention is then estimated as a normalized distribution of this energy:

$$a_t = softmax(e_t) \tag{1.3}$$

Once the attention is calculated, each individual attention value a^j_t is used to weight the representation h_j of each word, allowing the model to create a contextual representation c_t that focus on specific words at each step.

$$c_t = \sum_j a^j_t h_j \tag{1.4}$$

The context vector is concatenated with the decoder state and passes through a linear and a softmax layer to compute the probability of generating the output word $p_g(x^{'}_t)$:

$$P_g(x^{'}_t) = softmax(W^{'}(W[s_t, c_t] + b) + b^{'}) \tag{1.5}$$

where $W^{'}$, W, b, and $b^{'}$ are learnable parameters. We finally use a copy mechanism as presented in the *Pointer Generator Model* (PGN) (See et al., 2017) to consider Out-Of-Vocabulary words. The new probability of generating the output word $x^{'}_t$ becomes:

$$p_{gen} = \sigma(W^\top_{hgen} c_t + W_{Sgen} s_t + W_x x^{'}_{t-1} + b_{pgen}) \tag{1.6}$$

$$P(x_t^{'}) = p_{gen} \times P_g(x_t^{'}) + (1 - p_{gen}) \times \sum_{j:x_j^{'}=x_t^{'}} (a_t^j) \tag{1.7}$$

where σ is the sigmoïd function, W_{hgen}, W_{Sgen}, W_x, and b_{pgen} are learnable parameters. The model is trained with a standard negative log likelihood loss to optimize the generation probability distribution $P(X_i^{'})$ of the predicted document $T_i^{'}$, based on the reconstruction of the original input set X_i.

1.3.2 Informative constraint model

In the context of unsupervised single document summarization, the output should respect a length constraint represented by compression ratio $\alpha < 1$. If we directly apply this constraint to the Seq2Seq autoencoder, the model would simply produce the input αN_i first words of the input. Therefore, we need to implement an additional constraint to identify relevant information to preserve in the summary. Several methods using denoising, topic information, or word probability (Févry and Phang, 2018; Baziotis et al., 2019; Malireddy et al., 2020) are effective in constraining the model to produce a shorter output that respects this specific information. The well-known TFIDF metric emphasizes the relative contribution of terms within a text by counting their occurrences and their dispersion throughout the documents. This metric brings forward the specific information of a document and has proven to be efficient in many natural language processing tasks, especially for news stories that include events with salient features. In our approach, we aim at summarizing the current iteration text T_i which can be expressed as a set of words $X_i = \{x_1, x_2, ..., x_j, ..., x_{N_i}\}$ of size N_i. Based on the training dataset, we pretrained a *TFIDF* model, thus attributing a score for each word composing T_i. To create an information constraint, we create a second input for the model where we remove $1 - \alpha$ of the terms from T_i, those with the lowest *TFIDF* score, while preserving its token order. We call this new input the information constraint vector $TFIDF_i = \{x_1, ...x_j, ..., x_{M_i}\}$, where j still corresponds to the index associated with the original position of the term x in T_i, and $M_i = \alpha * N_i$ is the new size of this set which is set to the desired summary's length. For example, if we define a sentence $S =$ "The city emergency service confirmed a train accident", vectorized as $T_S = \{x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$. Depending on the pretrained TFIDF model, assuming a ratio $\alpha = 0.4$, a possible constraint sentence could be “emergency confirmed accident”. Therefore, we would have $TFIDF_S = \{x_2, x_4, x_7\}$. When testing our model, we replace each Out-Of-Vocabulary word by an unknown token $< UNK >$ indexed in the vocabulary. We specify the value of these tokens to the mean *TFIDF* of all the sentence terms to be able to include these terms in the summary. The model then must learn to encode and decode the two texts simultaneously, thus learning a language model and a vector implicitly defining the size of the summary and the material to preserve. Moreover, to account for text consistency and update information, we add two major modifications to the CM. Following the principle introduced by Luhn (Luhn, 1958), we apply a contextual window of size 2 and iteratively go through our text T_i. For each word belonging to $TFIDF_i$, neighbors' scores are multiplied by 1.25. We chose this value empirically after multiple tests because it promotes grammatical contextual words and preserves text coherence while retaining a enough of the original *TFIDF* top score words. These new scores thus increase the local coherence of the selected terms to form a new set $TFIDF_i^C$. Finally, to update information through each news iteration, we also consider the frequency of words present in the text generated $T_{i-1}^{'}$ by the model at the previous iteration $i-1$ to positively or negatively weight the score. The final CM is thus defined as follows:

$$TFIDF_i^u = \max_{w \in T_i} \sum_{j=1}^{\alpha N} \begin{cases} TFIDF_i^C(w_j) & \text{if } i = 0, \\ TFIDF_i^C(w_j) + \lambda\beta \times TF_{T_{i-1}^{'}}(w_j) & \text{if } i \geq 1 \end{cases} \tag{1.8}$$

where $TFIDF_i^u$ is the final set composed by the best scoring words for the summary; β is a parameter compensating the frequency of update terms and set as the mean *IDF* score of all terms, here 0.7; and λ is our consistency / novelty parameter that we will discuss later in the section 1.3.5.

1.3.3 Dual encoder

The second modification we made to the base model to account for update information is to input the representation of the previous produced text $T_{i-1}^{'}$ when predicting $T_i^{'}$. More specifically, if the current text to summarize T_i is the set of words $X_i = \{x_1, ..., x_j, ..., x_{N_i}\}$, the previous generated text $T_{i-1}^{'}$ can also be defined as the set $X_{i-1}^{'} = \{x_1^{'}, ..., x_j^{'}, ..., x_{M_{i-1}}^{'}\}$, where M_{i-1} is the size of the collection of terms generated at the previous iteration. At first iteration $i = 0$, we input an empty set for the summary. Therefore, the model acts as the regular autoencoder introduced section 1.3.1. Then, when $i \geq 1$ we provide both text T_i and $T_{i-1}^{'}$ to the encoder to obtain the document representation D_i, previous representation $D_{i-1}^{'}$, and the two encoding hidden states $h_1, ..., h_j, ..., h_{N_i}$ and $h_1^{'}, ..., h_j^{'}, ..., h_{M_{i-1}}^{'}$. We concatenate that information to obtain the final vector and set:

$$D_{T_i} = [D_i; D_{i-1}^{'}] \tag{1.9}$$

$$H_i = \{h_1, ..., h_{N_i}, h_1^{'}, ..., h_{M_{i-1}}^{'}\} \tag{1.10}$$

The new concatenated representation D_{T_i} and all encoding hidden states are provided to the decoder to estimate states and context vectors at each step t. It thus allows the decoder to focus on words from both the current text and previous text when generating $T_i^{'}$.

1.3.4 Loss function

Once the information constraint and the update are provided as additional input to the model, we need to set up an objective function that takes them into account. The purpose here is therefore twofold since the model must learn to respect the information in the CM and to follow a LM, allowing it to reconstruct the original texts properly. The goal for the CM is to predict a sequence $TFIDF_i^{'u} = x_1^{'}, x_2^{'}, ..., x_{\alpha N}^{'}$ such that we minimize the loss reconstruction related to the CM, which is defined by the cross entropy between this predicted vector and the original $TFIDF_i^u$:

$$\mathscr{L}_{TFIDF}(\theta) = \sum_{x \in TFIDF_i^u} \log p(x^{'}|x;\theta) \tag{1.11}$$

where θ are the model parameters. To ensure that the result generated by the model $TFIDF^{'u}$ following the information constraint remains consistent with the input texts T_i, we add a similarity loss between their representation. As in the case of (Chu and Liu, 2019), we re-encode the result $TFIDF^{'u}$ to obtain document representation $D^{'}_{TFIDF^{'u}}$, and we measure its cosine similarity with the initial D_{T_i}:

$$\mathscr{L}_{COS} = -(1 + cosine\_sim(D^{'}_{TFIDF^{'u}}, D_{T_i}))/2 \tag{1.12}$$

Since the cosine similarity can vary from -1 for opposite vectors to 1 for similar ones, we modify the loss in order to obtain normalize values from 0 to 1. This modification pushes the model to obtain a maximum similarity between the two representations. As for the LM, it is possible to increase the lexical and semantic diversity produced by a Seq2Seq model by adding a redundancy condition in its objective function (Li et al., 2016). In the case of update summarization, we thus train the model to account for the content of the preceding text it has generated. Therefore, the objective function becomes:

$$\mathscr{L}_{LM}(\theta) = \sum_{x \in T_i} \log p(x^{'}|x;\theta) + \lambda \sum_{x_{prev} \in T^{'}_{i-1}} \log p(x^{'}|x_{prev};\theta) \tag{1.13}$$

where $\lambda \in [-1, 1]$ is a control parameter that makes it possible to consider the update information. Once again, we will further demonstrate and discuss the influence of this control parameter in the generation of summaries by the model in the evaluation section 1.5. The final objective of the model is therefore to minimize the total loss:

$$\mathscr{L}_{TOT} = \mathscr{L}_{TFIDF} + \mathscr{L}_{LM} + \mathscr{L}_{COS} \tag{1.14}$$

1.3.5 Generation of novel or consistent summaries

During training, the lambda parameter lets us control information in the CM and enrich the LM in such a way that we can alternate between novelty and consistency between iterations.

- When $\lambda < 0$, equation (1.8) shows that the TFIDF scores will be penalized by the words occurring in the previously generated text. Then, regarding the loss function as defined in (1.13), we have the equivalent of the MMR approach. In this configuration the model attempts to minimize the likelihood of the previous summary. This is the analog of penalizing the similarity with the term distribution of the previous material. However, this adversarial approach can lead to some instability when $\lambda \leq -0.5$.
- When $\lambda > 0$, equation (1.8) once again shows that the TFIDF scores will be increased when words from previously generated text are repeated in the new summaries. Moreover, regarding the loss function in (1.13), the model tries to optimize the likelihood of both current and previous texts, thus preferring to include already seen content. In this configuration, it will enforce consistency between each produced update.

At each step, we provide the texts T_i, $T_i^{'}$, and the updated constraint representations $TFIDF_i^u$ as input to the decoder to respectively output a probability $P(x_{T_i}^{'})$ for the LM and $P(x_{TFIDF_i^u}^{'})$ for the CM. The summary is produced by maximizing the probability of the sequence of words such that:

$$\begin{gathered} P(x \in S_i;\theta) = \max_{x \in V}[P_{T_i}(x^{'};\theta);P_{TFIDF^u}(x^{'};\theta)] \\ \text{w.r.t } len(S_i) \leq L_S \end{gathered} \tag{1.15}$$

where V is the entire training vocabulary, and L_S is the maximum expected size of the summary S_i for the input text T_i. L_S is set such that $L_S = \alpha N_i$, where α is the compression ratio of the model as defined previously for the CM. At each step, the RNN decoder uses the previous generated word as an input for both models. This shared input makes both models start from the same input, while this competing approach allows us to select the most appropriate term with respect to either the information relevance of the CM or the coherence of the LM. The overall summary is produced with beam search decoding strategy improving the approximation of equation 1.15.

1.4 Experimental setup

1.4.1 Dataset

To evaluate our approach of update summarization of short simple documents, we use the 2013, 2014 and 2015 sections of the TREC track on temporal summarization from the KBA Stream Corpus (Frank et al., 2012). The dataset is composed of a set of documents answering a query on a specific event, created using hourly crawls. The task normally consists of extracting representative elements to constitute a summary of each update. These summaries are then compared to reference nuggets manually extracted from the same pool by expert assessors. We note in the dataset that many pairs (e.g. of news, nuggets) come from the same source, and are identifiable through the IDs of the documents provided in the dataset. Therefore, the dataset can be modified to effectively track the information issued by a single origin over time. Once we merge all three datasets, we obtain 6,186 news story for 10,839 text pieces. For our task, a news story then consists of a time series of text, composed of potentially multiple iterations, emitted by a single source of information. We perform several processing operations on the dataset, such as cleaning up URLs, the document ids, or the encoding problems that appear. Moreover, we filtered the news stories with fewer than 5 words and more than 100 words to reduce the noise in the input data. For computational reasons, we also omitted stories longer than five iterations (or updates). After applying these filters, we obtained 5,614 document streams, 69% of which contain only 1 text and are thus not subject to any updates, 17% of which have 2 texts, and 12% of which have 3 or more texts. The average size of a document issued from a news story is 38 tokens, and 16 for their associated summaries. We randomly split the dataset with a proportion of 70%, 20%, and 10%. In so doing, we obtained a total of 6,126 examples for training, 1,450 for validation, and 864 for final model testing.

1.4.2 Evaluation metrics

We first use the F1-ROUGE metrics for ROUGE-1, ROUGE-2 and ROUGE-L (Lin, 2004), which are standard for document summary evaluation. They respectively assess word overlap, bigram overlap, and the longest common subsequence between our references and the summaries generated by our model. We also complement our study with an unsupervised automatic metric: SUPERT (Gao et al., 2020), which measures the similarity of content embedded in the summary with the input text. This choice was made to address two major defects that stem from the fact that references are originally created for sets of documents. First, the references can be noisy, providing information from other news sources, and second they can be longer than the input. In light of these challenges, we evaluated our model using SUPERT, which fits particularly well with unsupervised approaches of text summarization. However, these evaluation metrics only shed light on our model's content relevance. To remedy this issue, we completed the assessment of our model with a human evaluation procedure to consider the grammatical quality of our model. Finally, we conducted several other classic analyses, such as an ablation study, or a sensitivity analysis, and we designed experimental protocols for understanding the novelty and consistency behavior of our approach. The detail of each experiment is detailed in section 1.5. The purpose of these analyses is not to evaluate the quality of the model, but to help us achieve a clearer portrait of our model as a whole.

1.4.3 Implementation

For our experiments, our model uses the GloVe 100 dimensional pre-trained word embeddings (version: glove.6B.100d) (Pennington et al., 2014). Both the encoder and the decoder of the model are composed of a single bidirectional layer with a size of 512 hidden units. We opted against using a more advanced architecture based on transformers and rather adopted a simpler architecture to better respond to the objectives of this study, which is to focus on the λ parameter controlling the information. The simpler architecture makes it possible to reduce the number of parameters, which in turn accelerated and facilitated the learning process knowing that the modification could improve any model structure. While the transformer architecture could outperform our present model and propose a more useful model, our objective is to explain and understand how to bias information to address novelty or consistency. Therefore, simplicity reduces the risk that an architecture with too many parameters and too much capacity may implicitly capture consistency- or novelty-related information, thus reducing the impact of our parameter and potentially distorting our results and analyses. We initialize the weight of the different layers via a Xavier uniform distribution (Glorot and Bengio, 2010), and we established the dropout of each layer at 0.2. To train the model, we conducted a Bayesian optimization of all hyperparameters based on the validation collection. Specifically, all hyperparameters are optimized by a defined-by-run strategy with the Optuna framework (Akiba et al., 2019). More specifically, we define a hyperparameter space and Optuna seeks the minimization of our objective function. We have then selected the best set of hyperparameters after 20 epochs to run our full implementation. Consequently, we train the model with the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-4}, a weight decay of 8^{-3} and a gradient clipping of 10. To allow the algorithm to learn to produce good quality texts, we first train the model without accounting for previous iterations. We start updating the CM and LM losses with updates at epoch 50 and we train the model for 80 epochs. Finally, we train the model with stochastic gradient descent using mini batches of dynamic size corresponding to the number of iterations of each news story. We then use the gradient accumulation technique to obtain a final equivalent batch size of 128. To generate the summaries, we define a compression ratio $\alpha = 0.4$ corresponding to the average compression rate in our training data. To improve further the output results, we apply the beam search method with a beam size set to five and an n-gram blocking (Paulus et al., 2017) set to avoid trigram repetitions. We implemented our model with the Pytorch library[5] version 1.8.1. and is available on GitHub [6]. The model was trained on a machine with an 8 Gb NVIDIA Tesla P4 graphics card and a 60 Gb 16-core processor.

1.5 Results and discussion

1.5.1 Model evaluation

To obtain a comprehensive view of our method, we compare our model with several baselines. Our first baseline consists in extracting the first 9 words of each document, reproducing the ROUGE principle. Essentially, this consists in extracting the first sentences, which are often used as reference for news stories (Févry and Phang, 2018; Nallapati et al., 2017). Then, we decide to compare our model with other recent abstractive approaches. For unsupervised approaches, since our dataset is composed of short texts (i.e. often one sentence in length), we chose sentence compression models as baselines. Specifically, we use the implementation of the denoising autoencoder 2-g shuf with Out-Of-Vocabulary words management as presented in (Févry and Phang, 2018) and the SEQ3 (Baziotis et al., 2019) consisting of two chained autoencoders that consider both topic information and sentence reconstruction in the compression process. We also report the performance of supervised models SummaRuNNer (Nallapati et al., 2017) and Pointer Generator Network (PGN) (See et al., 2017) to analyze their ability to capture updated information implicitly thanks to the provided references. Finally, we report the data from our model under three different configurations for Unsupervised Summarization for Updated Sentences (USUS). USUS_add_03 for $\lambda = 0.3$ and USUS_sub_03 for $\lambda = -0.3$. Table 1.1 presents the comparative results of our analyses for the ROUGE and SUPERT sores.

We observe that the supervised model Summarunner had the best results. The advantage of this model is that it can capture complex relationships between texts through human references. It is therefore difficult to compare it with our model. However, it is interesting to see that our model is the closest to SummaruNNer in performance and even exceeds

[5] https://pytorch.org/

[6] https://github.com/florentfettu/update-summarization-production

Table 1.1: Results for the TREC 2013-2015 dataset. Average results on the TREC 2013-2015 dataset for update summarization of short documents. **Bold numbers** only indicate the best results obtained using unsupervised methods, not statistical significance. Supervised results are presented for comparison and context.

Type	Methods	R-1	R-2	R-L	SUPERT
Baseline	First 9 words (F9W)	22.64	5.26	19.01	10.95
Supervised	SummaruNNer	26.78	20.12	26.34	13.14
	PGN	16.55	9.11	15.94	10.64
Unsupervised	2-g shuf Denoising AE	16.73	**3.16**	14.13	11.98
	SEQ^3 (Full)	15.67	0.835	11.26	13.24
	USUS_sub_03	16.7	2,13	**14.24**	15.64
	USUS_add_03	16.1	2.34	13.99	15.98
	USUS_sub_08	**16.76**	2.74	14.68	15.53
	USUS_add_08	16.05	2.28	13.95	**16.02**

the PGN architecture. We hypothesize that the relatively poor performance of the PGN is mainly due to its sensitivity to noise in the standard references, especially when the model learns the copying mechanism, whereas SummaRuNNer rather employs a word dropout technique at input level, making it more resilient to this phenomenon. The F9W baseline, which is created by extracting the 9 first words of each input sentence, is the second-best performing model on the F1 ROUGE score. However, we argue that this score is artificially increased by the summaries' size rather by content salience alone. First, we note that 28% of our input texts have fewer than 9 words after removing stop words. Therefore, the whole input acts as generated summary when computing the score with the reference. The score then reflects the correlation between input and gold nuggets. This hypothesis is corroborated by the SUPERT metrics where the F9W is outperformed by most models showing that it embeds less salient information from the document. Regarding our model, all our configurations (including our baseline that does not consider information update ($\lambda = 0$)) fairly compared with other unsupervised sentence compression algorithms for the ROUGE metrics. However, the contribution of our approach is clearly evidenced by the SUPERT metric analysis, where our model outperforms the other unsupervised methods. This once again demonstrates the usefulness of TFIDF for embedding salient information from a news event.

It is important to emphasize here that some limitations exists when evaluating produced summaries with automatic metrics such as ROUGE and SUPERT, especially when the objective is to gauge their quality. Following the work in (Févry and Phang, 2018), we conducted a human evaluation of our generated summaries to judge their fluency and the nature of the embedded content. We asked 5 native English speakers to assess the results generated from four models: USUS_sub_03, USUS_add_03, SEQ^3 (Full) (Baziotis et al., 2019), and 2-g shuf (Févry and Phang, 2018). We complement the results with the human references to have a fair baseline to compare the models. Each reviewer received a file containing the 5 shuffled summaries associated to 50 randomly selected texts from our test dataset. The evaluators were then instructed to consider two criteria when assessing the produced texts. The first criterion refers to the coherency of the text, which consisted in the summary only. The second criterion refers to the information quality embedded from the source. Here, the file included the original text accompanying the summaries. The evaluators were instructed to rate the summaries on a scale of 1 to 5 for each criterion. Table 1.2 presents the average evaluation of the summaries' coherence and content. A text receiving a score of 1 in both columns indicates that the summary has poor grammar and encompasses little of the original content.

Table 1.2: Human evaluation. Mean scores for 5 native English evaluators.

Methods	Coherence	Content	Redundancy	Consistency
Human references	4.4	3.7	2.16	3.92
2-g shuf	**2.83**	2.74	2.34	2.72
SEQ^3 (Full)	2.71	2.74	2.67	2.74
USUS_add_03	2.63	**3.22**	2.37	**3.27**
USUS_sub_03	2.69	3.13	**1.98**	2.96

The results corroborate our initial analyses, in particular the SUPERT score. According to our evaluators, our two models obtain the most relevant results when compared to the baselines. However, as we expected, the semi-extractive capability of our model creates a significant loss of grammaticality for our generated sentences when compared with human references. However, our method's results remain equivalent to other unsupervised abstractive methods. These results can be explained largely due to the combination of the LM, which acted as intended, and the modification of our TFIDF constraint to increase the score of neighboring terms to favor grammatical summaries.

As a result of this study, we observed another emerging trend. Our model USUS_sub_03 performs better on ROUGE while, regarding human evaluations and SUPERT score, USUS_add_03 encompasses more information from the source. As such, our two versions must produce text with different content. Our hypothesis is that USUS_sub_03 favors term novelty between two iterations, while USUS_add_03 enforces consistency between texts. However, it is not possible to draw firm conclusions from these results at this time given the study's limitations; that is, it is not possible to adequately evaluate text quality using only the scores obtained through ROUGE, SUPERT and human evaluators. Indeed, the dataset constitution, where gold references were issued from multiple documents, created some noise. The wrong associations between source texts and summaries may artificially increase the results of our novelty model since the summary may have no relation to its input, and therefore no relation to the previous iteration. In the case of SUPERT, the fact that the texts come from individual sources probably strengthens the impression of consistency between two updates. These factors might explain the performance differences of our two algorithms. To address these issues, we also analyzed the novelty and consistency of the generated results with automatic metrics and human evaluation.

1.5.2 Novelty and consistency evaluation

To demonstrate the influence of the λ value on the embedded information in the summaries and the novelty/consistency capacities of our model, we set up two automatic metrics to characterize the iterative outputs and the terms employed. The first metric reports the proportion of words that are reused from the source T_i in the summary S_i and that are not present in the previous iteration S_{i-1}. Our second metric measures the ratio of the number of terms in common between S_i and S_{i-1} compared to the same number for T_i and T_{i-1}. When we push for novelty, we expect the information reused from the source to increase while the information from the previous iteration to decrease, and vice versa when we push to produce consistent texts. We then proceeded to a sensitivity analysis of our model with several lambda values and the results for update summaries, where $i > 0$, are reported in the Table 1.3.

Table 1.3: Analysis of the composition of update summaries.

Methods	% re-use of new terms from T_i	% re-use of terms from S_{i-1}
2-g shuf Denoising AE	0.33	0.958
SEQ^3 (Full)	0.144	0.936
USUS_sub_03	0.819	0.562
USUS_add_03	0.801	0.837
USUS_sub_05	0.874	0.478
USUS_add_05	0.747	1.161
USUS_sub_08	0.914	0.152
USUS_add_08	0.73	1.242

These results highlight two interesting phenomena. The lower values for the direct re-use of source terms obtained in the comparative models (Févry and Phang, 2018; Baziotis et al., 2019) illustrate their greater capacity of abstraction, but also shows a higher incidence of hallucinations of words not found in the original documents. As for re-using terms coming from the previous summary, the results, close to 1, exhibit that the distributions of terms between two iterative summaries and two iterative input texts are respected and are thus not considered by those algorithms. Our approach also emphases different characteristics for the re-using terms. As the λ value increases, the re-use of terms from the previous iteration increases while the terms from the current source decreases. The opposite phenomenon is observed when we decrease the lambda value. It demonstrates the impact of the λ parameter on the generated output. If there is an increase in the rate of terms re-used from the previous iterations, the model should increase the consistency of the text; if it decreases while favoring the use of terms in the current iteration, then the model should reinforce the novelty. This observation tends to follow the hypothesis that the novelty implementation avoids information overlap, which in turn reinforces the relevance of the results to a user (Goldstein et al., 2000), while consistency reinforces the understanding of the source information content (Barzilay and Elhadad, 1999).

We completed this study by instructing the human evaluators to consider two additional criteria when assessing the produced summaries. Using the same random sample of 50 abstracts, we gave them a file indicating the chronology of the summaries' updates for each event. The first criterion related to the update's redundancy with respect to the summary of the previous iteration. The second criterion concerned how closely the update summary was related to the given event described in the previous iteration. The evaluation was once again conducted using a scale of 1 to 5. Table 1.2 illustrates the results for different models and the human references. We observe that our USUS_sub_03 model with a negative value of λ produces texts with a lower redundancy than at baseline As for the USUS_add_03 model with a positive value of λ, we note that the human evaluators could more clearly gauge that the two iterations reported the same event. These results confirm that our two approaches can modify the content of the generated summaries.

We can further observe this difference with the example provided in Table 1.4, produced by our models USUS_sub_03 and USUS_add_03. These examples demonstrate the influence of the lambda parameter on the final production of summaries. Indeed, in the first iteration, we notice that the word "election" is re-used in the model that pushes for consistency while "mass protest" is favored by the novelty version. The observation is even more telling for the second instance, where the term "saturday" is employed by the consistent variant whereas it is non-existent in the original text of this iteration. Similarly, whereas "putin", a word that does not appear in the first iterations , is pushed very early in the sentence by the novelty model. This analysis of our produced summaries and both automatic and human metrics allow us to safely conclude that our proposed method is indeed able to promote the novelty or the consistency of information, depending on the choice of parameters.

Table 1.4: Example of generated texts. News stories on Russian elections. The event is updated 3 times. The example shows the summaries generated by the two models USUS_sub_03 and USUS_add_03. The table contains only filtered texts without stopwords and other grammatical terms to focus on information content. New or modified information that appears in subsequent iterations in the two models as a result of the parameter has been *italicized*.

	Original text	Summary $\lambda = -0.3$	Summary $\lambda = +0.3$
0	shot russian elections last sunday video one many examples alleged election fraud went viral started antigovernment protests russia	many russian *election* fraud went viral started antigovernment	many russian *election* fraud went viral started many
1	russian mass protests election results scheduled saturday 30000 people allowed gather moscow s bolotnaya square 11 cities russia also received official permits	russian *mass protests* scheduled saturday 30000 allowed gather cities	russian *election* results scheduled *saturday* 30000 allowed gather russia
2	anti putin activists promoting countrywide protests next saturday suggest shifting internationally hosted websites facebook opposed local social network v kontakte	promoting *putin activists* shifting internationally hosted websites kontakte	promoting *saturday suggest* shifting internationally hosted websites kontakte

1.5.3 Ablation Study

We conducted an ablation study on our consistent model with $\lambda = +0.3$ to further demonstrate the contribution of the different elements in our model. We place a particular focus on analyzing the implication of adding the summarization information in the dual encoder, the lambda parameters in both the language model and the CM. We first removed the summary in the encoder, thus depriving the system of that information. We observed that the model was unable to continue learning to reconstruct the text. Once it has to consider iterations in the loss function, the decoder experiences a gradient explosion because it does not have enough information to follow the update objective. This demonstrates the critical role of our dual encoder in the model but also its potential instability.

- When $\lambda = 0$ for both models, the embedded information stays relatively similar, with a SUPERT score of 15.9 and a ROUGE1 F1 score remaining at 16.01. However, However, the re-use rates of terms from T_i is 0.8 and S_{i-1} is 0.624, indicating that the model is no longer able to manage the re-use of terms from previous iterations. This shows that now we only aim to maximize the likelihood of the current text and the CM is equivalent to a simple TFIDF optimization. In this way, the approach is similar to existing methods for sentence compression.

- When $\lambda = 0$ for the CM only, we note the same behavior and performance as for the previous configuration where all values set to 0. This provides further evidences that the model relies heavily on the TFIDF metric to produce updated information.

- When $\lambda = 0$ for the LM only, it results in a difference between the two rates of term re-use in the summaries, namely, 0.723 for new terms from T_i and 1.234 for terms from S_{i-1}, which is equivalent to the highest rate observed for $\lambda = 0.8$. However, there is a significant drop in performance for ROUGE and SUPERT –almost 2 points each. This trend seems to confirm that without the LM, the model overfits the TFIDF CM and no longer produces as relevant and coherent outputs.

1.5.4 Conclusion on the results analysis

The automatic and human analyses demonstrate that the model obtains satisfactory results, which are either equivalent or superior to existing state-of-the-art models in terms of preserving information from the originals texts. Additionally, despite its semi-extractive capabilities, our model produces results that remain understandable and sufficiently coherent for those fluent in English. Furthermore, as evidenced by the experiments, this study shows that it is possible to instruct the model to generate summaries that prioritize either content novelty or content consistency. The ablation study confirms the importance of the lambda parameter in both language and constraint models to deal with iterative information. These results fully highlight the interest of methods that consider updating information where documents are temporally correlated, such as TREC news streams.

1.6 Conclusion to chapter 1

In this paper, we presented an unsupervised autoencoder method for semi-extractive document summarization. It relies on two competing models that either generate coherent text or control the information included in a summary. By defining explicit constraints and objective functions, we were able to introduce parameters that account for novelty and consistency of information through iterative and/or streams of texts. Therefore, the proposed approach addresses the new task of update sentence compression or short update summarization. As a result, the model outperforms state-of-the-art unsupervised abstractive sentence compression systems for specific datasets such as the TREC temporal track. The model also modulates the information present in the final output, making it more flexible and appropriate for meeting specific needs.

Of course, our approach comes with certain limitations that stem from the architecture used to implement the novelty and cohesiveness parameter. First, in terms of the novelty, we introduce an adversarial learning issue in the LM, can lead to gradient explosions, which in turn creations instability in the learning process. This restricts our method to small values of lambda for novelty, thus capping the impact on generated summaries. We obtained good results from initial tests with two lambda value, a small value for the LM and a large one for the CM, which opens the door to further study on stabilization techniques that are likely to improve the model's ability to address novelty. Due to our hardware limits but also with our desire to study explicit information control, we are aware of the shortcomings of our approach compared to recent architectures based on transformers. However, we believe that our findings can be applied to these models in order to increase their portability and their ability to be used for various tasks. In future work, we further hope to consider models with more capacity, especially generative variational autoencoders to understand the impact of our parameters on constraining latent representations. We also plan to investigate different possible solutions for our information constraints. Our current approach only considers shallow statistical metrics to emphasize text relevance. However, the addition of linguistic data such as dependency trees, lexical chains, or discourse-oriented features could both improve the grammaticality of the model and its capacity to identify consistent or novel content between two updates. Finally, it would be interesting to apply our model to the task of update or iterative sentence compression for long text summarization. Since two following sentences in a document have local information correlation, we expect that models accounting for novelty or consistency could allow automatic systems to efficiently reduce the size of such documents. Notwithstanding these limitations, and given the large potential for improvement, this study positions our model as a strong baseline for assessing the performance of upcoming update sentence compression and summarization tasks.

References

Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. (2009). Diversifying search results. In *Proceedings of the second ACM international conference on web search and data mining*, pages 5–14.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Allan, J., Gupta, R., and Khandelwal, V. (2001). Temporal summaries of new topics. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 10–18.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Barzilay, R. and Elhadad, M. (1999). Using lexical chains for text summarization. *Advances in automatic text summarization*, pages 111–121.

Baziotis, C., Androutsopoulos, I., Konstas, I., and Potamianos, A. (2019). SEQˆ3: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression. In *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)*, pages 673–681, Minneapolis, Minnesota. Association for Computational Linguistics.

Bysani, P. (2010). Detecting novelty in the context of progressive summarization. In *Proceedings of the NAACL HLT 2010 Student Research Workshop*, pages 13–18.

Carbonell, J. and Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Chu, E. and Liu, P. (2019). Meansum: a neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*, pages 1223–1232. PMLR.

Delort, J.-Y. and Alfonseca, E. (2012). Dualsum: a topic-model based approach for update summarization. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 214–223.

Fabbri, A., Li, I., She, T., Li, S., and Radev, D. (2019). Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings* *of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Févry, T. and Phang, J. (2018). Unsupervised sentence compression using denoising auto-encoders. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 413–422, Brussels, Belgium. Association for Computational Linguistics.

Frank, J. R., Kleiman-Weiner, M., Roberts, D. A., Niu, F., Zhang, C., Ré, C., and Soboroff, I. (2012). Building an entity-centric stream filtering test collection for trec 2012. Technical report, Massachusetts Inst of Tech Cambridge.

Gao, Y., Zhao, W., and Eger, S. (2020). SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Ghadimi, A. and Beigy, H. (2020). Deep submodular network: An application to multi-document summarization. *Expert Systems with Applications*, 152:113392.

Ghadimi, A. and Beigy, H. (2022). Hybrid multi-document summarization using pre-trained language models. *Expert Systems with Applications*, 192:116292.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Goldstein, J., Mittal, V. O., Carbonell, J. G., and Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. In *NAACL-ANLP 2000 Workshop: Automatic Summarization*.

Kedzie, C., McKeown, K., and Diaz, F. (2015). Predicting salient updates for disaster summarization. In *Proceedings of the 53rd Annual Meeting of the Association for Com-*

putational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1608–1617.

Kim, G. and Ko, Y. (2021). Effective fake news detection using graph and summarization techniques. *Pattern Recognition Letters*, 151:135–139.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (2016). A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Li, P., Ye, Q., Zhang, L., Yuan, L., Xu, X., and Shao, L. (2021). Exploring global diverse attention via pairwise temporal relation for video summarization. *Pattern Recognition*, 111:107677.

Lin, C., Ouyang, Z., Wang, X., Li, H., and Huang, Z. (2021). Preserve integrity in realtime event summarization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(3):1–29.

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

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

Malireddy, C., Maniar, T., and Shrivastava, M. (2020). Scar: sentence compression using autoencoders for reconstruction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 88–94.

McCreadie, R., Macdonald, C., and Ounis, I. (2014). Incremental update summarization: Adaptive sentence selection based on prevalence and novelty. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 301–310.

Mohammed, M. B. and Al-Hameed, W. (2021). Cohesive summary extraction from multi-document based on artificial neural network. In *2021 7th International Conference on Contemporary Information Technology and Mathematics (ICCITM)*, pages 81–87. IEEE.

Nallapati, R., Zhai, F., and Zhou, B. (2017). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Paulus, R., Xiong, C., and Socher, R. (2017). A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Prabhumoye, S., Quirk, C., and Galley, M. (2019). Towards content transfer through grounded text generation. In *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)*, pages 2622–2632, Minneapolis, Minnesota. Association for Computational Linguistics.

Rudra, K., Ganguly, N., Goyal, P., and Ghosh, S. (2018). Extracting and summarizing situational information from the twitter social media during disasters. *ACM Transactions on the Web (TWEB)*, 12(3):1–35.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

West, P., Holtzman, A., Buys, J., and Choi, Y. (2019). BottleSum: Unsupervised and self-supervised sentence summarization using the information bottleneck principle. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3752–3761, Hong Kong, China. Association for Computational Linguistics.

Zhao, B., Li, H., Lu, X., and Li, X. (2021). Reconstructive sequence-graph network for video summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2793–2801.

Zhao, B., Li, X., and Lu, X. (2018). Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7405–7414.

Zhu, W., Lu, J., Li, J., and Zhou, J. (2020). Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, 30:948–962.

Chapter 2

Objective and neutral summarization of customer reviews

Abstract

Opinion mining aims at detecting and extracting relevant information from large quantity of customer reviews. Automatic opinion summarization then seeks to create a consensual point of view often oriented toward the main sentiment of clients to render their experience. Although factual information is valuable for companies to understand what works or not in their products, no summarization approach has been conceived yet to convey objectivity and constructive feedback from customer reviews. To address this new issue, we propose an adversarial multi-task learning model for document summarization. Our algorithm combines an autoencoder for document summarization with a gradient reversal layer to learn agnostic representations to subjective and sentiment-based material. We assess and compare our method on the Amazon product review dataset where we introduce an original evaluation dataset for objective summarization. We further completed the analysis with neutrality and objectivity metrics. This study demonstrates that the generated summaries carry out both relevant and objective content, but also emphasize the importance of various processes and layers in multi-task learners to control effectively the information.

2.1 Introduction

Mining information in customers' opinions has always been crucial to understand the quality of a society or a commodity. These reviews provide excellent material useful to both potential consumers and product manufacturers. Specifically, from an organizational point of view, they are relevant to take decisions according to the analysis results about their products. For example, in the financial sector, the Net Promoter score and its associated comments have become one of the essential markers for judging the performance of an institution and letting it improve its services (Reichheld, 2006). Due to the subjective nature of sentiments, looking at only the opinion from a single person is usually insufficient. Therefore, summarization is exceedingly meaningful in this context to bring out a digest and actionable report from this data.

Automatic text summarization is the process of distilling the information contained in one or more documents to produce a reduced version that meets the need for a particular task or user. In the context of customers' opinions, a major issue resides in the input diversity, depending on the products or company concerned by the comments (Blitzer et al., 2007a). Unsupervised learning then becomes valuable because of the portability of such approaches, not requiring prohibitive amounts of labeled data to obtain performing models (Mihalcea and Tarau, 2004). The first methods aimed at selecting the most redundant content of the reviews (Ganesan et al., 2010). Recent advancements in deep learning with self-supervised techniques allow researchers to propose efficient abstractive solutions. The objective for these models is to create an average representation of the data then learn to decode it into realistic consensual output (Chu and Liu, 2019; Bražinskas et al., 2019). Whatever the approach, the principal content has always been considered as correlated to the main sentiment because it carries a credible input of people's experience. Therefore a lot of research has biased information selection toward the sentiment of the corpus and its extreme members (Pang and Lee, 2004; Hu and Liu, 2004; Cao et al., 2017). It is particularly hard to produce a unanimous summary due to the conflictual nature of individual opinions (Pang and Lee, 2004). Some recent works thus propose to restrain the scope by extracting aspect based information to gather precise sentiment for the summary (Angelidis and Lapata, 2018a). However, the question of its utility can still be questioned as it does not relay factual information but personal feelings (Liu et al., 2007).

Despite being significant for addressing concerns from customers who are very dissatisfied with certain aspects of a product and reacting immediately to prevent the situation from worsening, sentiment-oriented summaries can fail to offer appropriate and considerate feedback (Baron, 1993). Mainly because apprehending customers reality does not focus on the factual information that will target specific problems, and be congruent and descriptive, allowing it to plan further development (Whetten and Cameron, 2005). Moreover, this feedback might have negative effects by hurting staff and their motivation to remedy customer issues (Kipfelsberger et al., 2016). From this perspective, we believe a useful summary must provide a detailed view on the qualities and the shortcomings of the product or service enabling the company to learn from its mistake and take action accordingly. Predominantly, qualitative and quantitative outlooks have been extracted from subjective reviews through the scope of multi-document summarization (Liu et al., 2007) without regarding the possibility of capturing more neutral, objective and factual information. However, this compensation by the number does not solve certain problems related to subjectivity and sentiments. First, as we already mentioned, conflicting and opposed opinions are important challenges (Pang and Lee, 2004). The same authors further discuss the existence of a bias toward extremely positive experience because of how online platforms value such customer reviews . Authors in (Hu and Liu, 2004) also denote that subjective sentences tend to be longer than objective ones. These three properties thus induce an issue when assessing content relevancy due to misleading measure and interpretation of information redundancy. Finally, less fundamental but equally important point to consider is that subjective or extreme sentences often contain inappropriate messages such as abuse (Hu and Liu, 2004) that we definitely want to exclude from any summaries.

Therefore, in this article, we present a new model with a multi-objective model for unsupervised document summarization of product customer opinions to help organizations to get objective and constructive feedback. More specifically, we employ a self-supervised autoencoder to learn text representation and to decode them in a coherent output. We choose this method as it serves as a reference for unsupervised summarization. We then modify its functioning by adding a Gradient Reversal Layer (GRL) to classify sentiment associated to the review representations. This adversarial layer allows the model to influence encoded representations to ignore or prohibit sentiment and subjective information. During text generation, we can therefore use this biased encoding to produce our objective summaries. The contributions of this work are the following:

- We propose a novel approach and perspective to unsupervised customer review summarization focusing on including relevant objective content.
- We then propose a method to learn to perform this task. This model offer promising performances and can serve as a strong reference baseline for future studies.
- We demonstrate that GRL are efficient techniques for representation disentanglement to condition text generation.
- Finally, we bring forward a new evaluation dataset for Amazon product reviews composed of 200 human-generated objective summaries. We also provide complementary metrics to assess subjectivity and neutrality in automatically created output.

2.2 Related work

2.2.1 Multidocument Summarization for Opinion

In the context of customer reviews and opinion summarization, evaluation of relevant information relies heavily on information and feature redundancy. Classic extractive techniques such as TextRank (Mihalcea and Tarau, 2004) or more recently BERT and GPT-2 for extractive summarization (Miller, 2019) show good performances in this redundant context and are still nowadays used as a referring baseline for evaluating other methods. However, opinion-specific approaches have been suggested to handle this data. In their article (Ganesan et al., 2010), the authors introduce Opinosis, a semi-abstractive system based on a term co-occurrence graph to extract original and highly recurrent sequences. With the possibilities offered by deep learning, several autoencoder models have generated abstractive summaries depicting a consensus representation of the customers' point of view (Chu and Liu, 2019; Bražinskas et al., 2019). Product features are a crucial aspect of opinion definition, therefore authors in (Angelidis and Lapata, 2018b) create aspects-based representations with a partial autoencoder and devise an optimization function to select opinion that leverages their coverage. OpinionDigest (Suhara et al., 2020) is another method that clusters topically related reviews and employs a ranking algorithm to increase diversity in the output. Finally, (Amplayo et al., 2021) have introduced an interesting hybrid procedure that once again clusters opinions and extracts sentences to produce a summary predicated either on popular or specific aspects. Regarding abstractive summarization, authors in (Coavoux et al., 2019) combines Meansum (Chu and Liu, 2019) with a clustering algorithm to conceive latent representation for each group and form a text that maximizes input coverage. In the context of opinion summarization, the sentiment associated with this feature is also an essential factor to determine how important the aspect described represents the customer experience. As early as 2004, the authors of (Hu and Liu, 2004) proposed a system that predicts the sentiment paired with attributes to produce sentiment oriented summaries. This notion has been taken up in articles such as (Lovinger et al., 2019) where their model, Gist, attempts to create a typical review consisting of a few key sentences that will capture the main sentiment. ASMUS is another recent model (Abdi et al., 2019a) that performs a sentiment analysis separating positive and negative opinions to generate a distinct summary for each dimension. In the context of the deep learning paradigm, (Angelidis and Lapata, 2018b) proposed a mutli-task learning model for both extracting attributes and predicting their sentiment. They then create heuristics based on this information in an optimization framework to create the summary. In (Pecar, 2018), the author devises abstractive summaries predicated on positive and negative data, and a structured analysis of the product features. However, if the coherence and relevance of the reviews contribute to the outputs texts quality, it is not obviously the case for subjectivity

(Liu et al., 2007). On the contrary, it may participate in developing a false positive bias, or to include misleading, extreme, or inappropriate material in the final summary. Based on these observations, our approach differs from previous work because we consider that the relevance of the information is represented by its salience and redundancy but also by its neutrality or at least its stronger partiality to obtain constructive criticism of products.

2.2.2 Multitask and adversarial learning

The main goal of multi-task learning (MTL) is to improve a model performance and generalization ability on a task via the optimization of a second related-task objective (Caruana, 1997). This paradigm has proven particularly useful for NLP, especially with the emergence of general pre-trained models, and the need for huge quantities of training data. For our purposes, MTL has also been proven useful for documenting summarization and sentiment analysis. In (Cao et al., 2017), the authors demonstrate that adding a topic classification makes it possible for a model to generate qualitative and specific summaries with respect to the topic covered in the input documents. Regarding sentiment analysis, several MTL models have been designed to improve the task. One of the recent state-of-the-art models is bolstered by topic modeling since the terms can have a different sentiment polarity depending on the topic addressed (Gui et al., 2020). Moreover, Domain Adaptation (DA) has become one of the central themes of sentiment analysis and other NLP tasks. One of the techniques for this task is to use gradient reversal layers (GRL) (Ganin and Lempitsky, 2015). The MTL objective here becomes adversarial since we multiply the gradients by a negative factor such that we don't learn to predict the original domain of the data. We then obtain domain agnostic representations to perform the initial task. This is especially useful for sentiment analysis since pivotal features exist and can be domain independent (Blitzer et al., 2007b). This has been recently demonstrated for example in the case of filtering tweets expressing a sense of emergency (Krishnan et al., 2020), or purely for the domain shift between different Amazon product categories (Tang et al., 2021). The authors also obtain strong performances partly due to the decorrelation between the representation of dependent and domain-independent features in the model. Our approach mimics this idea to develop an MTL summarization model where the primary objective is to learn a language model by reconstructing individual reviews. We also consider some gradient reversal layers. However we transform the domain to become the reviews' rating to emphasize sentiment agnostic information. While this approach has been widely adopted for domain adaptation to enhance sentiment analysis (Ahmet and Abdullah, 2020) and to improve the generalization of classification tasks (Seng and Wu, 2023), using this technique for content generation is quite novel and has shown promising results for images (Havaei et al., 2021). We follow up on this concept by applying GRL for disentangling representations in a text generative context.

2.2.3 Subjectivity and sentiment

Customer reviews are usually composed of opinions and facts. We consider in this case that subjective text fragments convey important opinionated information while objective texts state facts that are less relevant since it doesn't express the customer experience (Chaturvedi et al., 2018a). Although subjectivity and sentiment analysis are related tasks, one distinguishes them because subjectivity includes the degree of an opinion sentiment and the attitude of the emitter (Medhat et al., 2014). In the context of opinion mining or summarization, many methods start by predicting subjective text segments to only input this information to create a summary from these facts (Angelidis and Lapata, 2018a). However, to provide a concise information that enables a company to explain why certain aspects have been criticized and how to propose improvements, we must perform the task on many opinions to compensate for the individuality aspect of subjective material (Ganesan et al., 2010). It's therefore essential to propose tempered summaries representing consensus among the largest number (Pang et al., 2008). Moreover, a neutral or less subjective opinion can still contain relevant information. Indeed, neutrality means that sentences can be contrasted by both positive and negative features (Colhon et al., 2017; Tsytsarau and Palpanas, 2012). They can also have modifiers that contribute very little to the sentiment but play on the subjectivity (Chaturvedi et al., 2018b). Or finally, the emitter intention is to appear more reasonable when he transmits his point of view. Unquestionably, neutral terms such as “so-so” or “mediocre” are perfect to communicate one opinion while decreasing the polarity of terms such as “bad” or “horrible” (Pang et al., 2008). Authors in (Wilson et al., 2009) perfectly re-transcript this principle by setting up a lexicon of gradations of terms subjectivity, distinguishing strong subjective with extreme polarity that can sometimes be inappropriate and weak ones [1]. Therefore, in our approach we use an abstract model based on an autoencoder to capture information consensus between reviews via redundancy, while neutralizing, and reducing the impact of strong sentiment without losing the generated summary subjective point of view.

2.3 Proposed Model

This section presents the general architecture behind our approach. Following the system introduced in (Chu and Liu, 2019), the vanilla model relies on a two-step-learning process. In the first phase the model learns a language model based on the customer reviews. In the second stage we train the summarizer by employing the mean representation of those reviews. Our main contribution, is the adversarial multi-task objective employed in this first step. Indeed, we add the gradient reversal layer to create sentiment agnostic representations, that will be used in the latter training phase and during generation.

2.3.1 Model background

In this article we use an encoder-decoder Seq2Seq architecture to model our documents. The corpus is composed by an ensemble of customer reviews on different products and services. The total vocabulary of the corpus is denoted V. Let’s define a batch of M customer reviews regarding a specific product as $\{R_1, ..., R_i, ..., R_M\}$ used to train our model. Each review R_i in M is composed of a set of words $X = X_1, ..., X_i, ..., X_N$. We

[1] `https://mpqa.cs.pitt.edu/lexicons/subj_lexicon/`

transform each input review with an embedding layer to obtain a first sentence matrix $E \in \mathbb{R}^{d*N}$, where d is the embedding space dimension. The sentence embedding matrix is then fed to a bidirectional Gated Recurrent Unit (GRU) (Cho et al., 2014), producing a sentence representation H_m, which is the last hidden state output of the layer. When decoding, we use H_m to initialize the initial hidden state S_0 of another simple GRU layer. At each step t, we feed the decoder with the preceding state S_{t-1}, the preceding generated word x'_{t-1}, and a contextual layer $C = H_m$:

$$s_t = GRU(s_{t-1}, x'_{t-1}, C) \tag{2.1}$$

At each step, we also input the real preceding word X_{t-1} using a teacher forcing method. Finally, the context vector C, the current hidden state S_t, and the current input embedding are provided to a linear then a softmax layer to compute the probability of generating the output word $p_g(x'_t)$ through our vocabulary V:

$$P_g(x'_t) = softmax(W'(W[s_t, c_t] + b) + b') \tag{2.2}$$

where W', W, b, and b' are learnable parameters. The model is trained with a standard negative log likelihood loss to optimize the generation probability distribution $P(X'_i)$ of the predicted review R'_m, based on the reconstruction of the original input R_m:

$$\mathscr{L}_{gen}(\theta) = \sum_{x \in R_m} \log p(x'|x; \theta) \tag{2.3}$$

2.3.2 Adversarial sentiment analysis

This section presents the modification we bring to the vanilla model to perform objective summarization. As seen in figure 2.1, we add the gradient reversal layer to create sentiment agnostic representations, that will be used in the latter training phase and during generation.

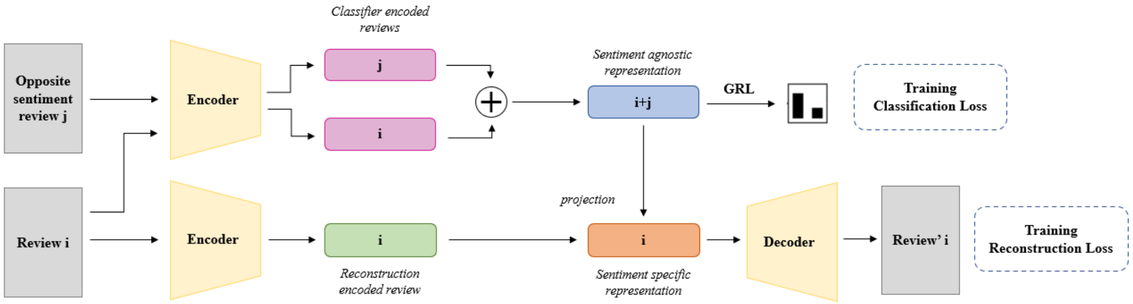

Classifier encoded reviews
Opposite sentiment review j
Encoder
j
i
Sentiment agnostic representation
i+j
GRL
Training Classification Loss
Review i
Encoder
i
Reconstruction encoded review
projection
i
Sentiment specific representation
Decoder
Review' i
Training Reconstruction Loss

Figure 2.1: Training the language model and sentiment bias representation. The figure presents introduces the training stage where the model learns language model and the sentiment bias.

Gradient reversal training

During the training of the language model, we add a second objective to predict the sentiment of each review. Each review R_i is associated with a sentiment S_i, which can be either positive or a negative. First, R_i is fed to another bidirectional GRU encoder to obtain the hidden vector H_i^{cls} for sentiment classification. Our objective being to produce sentiment agnostic representations, we make several modifications to the classification model. We concatenate H_i^{cls} with H_j^{cls}, which is a representation of R_j, a review of the same batch that have an opposite sentiment. Here is an example of an association that could be created by the model for some headsets: R_i = "I liked the headset so much that I also bought one from a friend's birthday.", and R_j = "The product arrived was sleek in its appearance, yet that's where the excitement ended. Call volume was low and the amount of static". This new concatenated hidden vector H_{ij}^{cls} is then fed to a Gradient Reversal Layer (GRL) (Ganin and Lempitsky, 2015) to obtain $\check{H}_{ij}$. As discussed in (Tang et al., 2021), the concatenation of reviews is essential to ensure that the model capture shared features from the two domains. Indeed, if we only use R_i as an input for the GRL, the model might associate agnostic words to the opposite sentiment. In our example, terms like "so" ", "much", or "the" are the best variables for predicting non-positive sentiment, thus being potentially paired with negativity. By having the second sentence, words like "ended" or "yet" become essential variables and the neutral words remain shared features. This alignment will be crucial to creating an orthogonal representation of this space, thus generating independence with journal sentiments. If the assignment of terms to a positive or negative aspect is balanced through learning with an average learned during learning only, then they will not be aligned with this independent space. The best way for them to remain completely independent is to allow the model to rely on truly negative or positive terms from the opposite sentiment review paired with our text. Therefore, the required condition for guaranteeing this alignment is to ensure that the reviews are positive or negative and paired together for each training point. As we will detail in 2.5.2, the important is to properly aligned this representation with the sentiment space, to ensure that the following projections convey the agnostic sentiment content. Finally, the classifier uses $\check{H}_{ij}$ to predict the sentiment S_i through a linear and a softmax layers:

$$P(S_i^{'}) = softmax(W[\check{H}_{ij}] + b) \tag{2.4}$$

The loss then follows:

$$\mathscr{L}_{cls}(\theta) = \sum_{i=1}^{M} \log p(S_i^{'}|S_i;\theta) \tag{2.5}$$

For the remaining of the article we simplify the notation of $\check{H}_{ij}$ to $\check{H}_i$ since the representation is biased toward S_i. We repeat this process of selecting opposite sentiment reviews and create these representations $\check{H}$ for each review in our batch.

Sentiment biased Language Modeling

As specified in section 1.3.1, the base model applies the encoded reviews from the GRU layer H for text reconstruction. If we directly train the language model with H, we do not fully use the potential of our adversarial multi-task approach because the model learn the reconstruction and the agnostic representation, respectively H and $\check{H}$, independently. To ensure the representation encapsulate sentiment information when recreating reviews, we follow the idea presented in (Tang et al., 2021), by projecting H on $\check{H}$ space :

$$\tilde{H} = \frac{H \cdot \check{H}}{|\check{H}|} \cdot \frac{\check{H}}{|\check{H}|} \tag{2.6}$$

where $|\check{H}|$ represents the norm of the vector $\check{H}$. $\tilde{H}$ then represents the projection of our reconstruction representation H on $\check{H}$. We then use $\tilde{H}$ as the initialization S_0 and the context C hidden layer for the decoder and in equation 2.2. While the projection constrains the reconstruction to be sentiment biased, the language model loss $\mathscr{L}_{gen}$ then encourage $\tilde{H}$ to encapsulate the maximum of relevant information for review prediction. We finally combine both losses to train the model such that we minimize the total loss:

$$\mathscr{L}_{tot} = \mathscr{L}_{gen} + \mathscr{L}_{cls} \tag{2.7}$$

With this projection mechanism we obtain a true joint learning for our multi-task autoencoder model.

2.3.3 Summarization phase

For the summary generation, following the procedure introduce in (Chu and Liu, 2019), we create a mean representation of all review to be summarized H_{mean}. For decoding, we fed H_{mean} to our decoder, as the initialization of the hidden state and as context vector, to generate a sentence that will be considered as our summary. As recommended, since we do not have ground truth summaries to train the model, we use a Straight GumbelSoftmax implementation (Jang et al., 2016) to alleviate exposure bias of teacher-forcing during learning. We also re-encode the summary generated to obtain H_{summ}, and we aim to reduce the cosine similarity with the set of individual review representations. This procedure is mainly used for making sure that the summary is in the same word vector space as the input reviews. However, as seen in figure 2.2, we modify this step to align the vector with the agnostic sentiment space.

As $\tilde{H}$ serves as an encoding of the review in the combined sentiment and information space, we can project H on the orthogonal space associated to $H - \tilde{H}$:

$$\hat{H} = \frac{H \cdot (H - \tilde{H})}{|H - \tilde{H}|} \cdot \frac{H - \tilde{H}}{|H - \tilde{H}|} \tag{2.8}$$

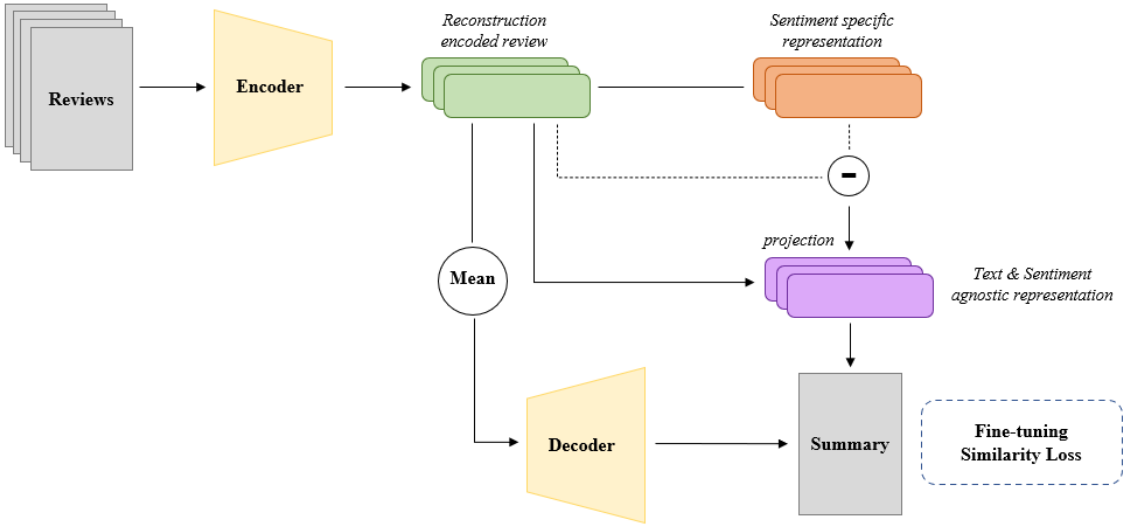

Reviews
Encoder
Reconstruction encoded review
Sentiment specific representation
projection
Text & Sentiment agnostic representation
Mean
Decoder
Summary
Fine-tuning Similarity Loss

Figure 2.2: Fine-tuning of the model. The figure presents the fine-tuning step where we bias again the summary towards neutral/objective representations.

This new review representation $\hat{H}$ thus encapsulate the maximum information of H learned for predicting text reviews and is orthogonal to our sentiment specific space thus creating an agnostic representation. We then use $\hat{H}$ to fine-tuning our model by reducing its cosine similarity with the decoded summary H_{summ}:

$$\mathscr{L}_{sim} = 1 - \sum_{i=1}^{M} cosine\_sim(\hat{H}_i, H_{summ}) \tag{2.9}$$

This fine-tuning procedure thus contribute to make sure that the model, while generating coherent summary, also produce a neutral or objective output.

2.3.4 Generation phase

Once the model is fully trained, we generate a review acting as the final summary R_{summ} of a batch of reviews $M = \{R_1, R_2, ..., R_M\}$. To this end, we pass each review to our two encoders to obtain $H_{(1...M)}$ and $\check{H}_{(1...M)}$. Since we have trained the model to capture information and fine-tuned it toward sentiment agnostic predictions, we can simply initialize the trained decoder GRU hidden state S_0 as H_{mean} and the context vector as $\hat{H}_{mean}$. The probability of generating the output word $P_g(x_t^{'})$ through equation 2.2 thus becomes:

$$P_g(x_t^{'}) = softmax(W^{'}(W[s_t, \hat{H}_{mean}] + b) + b^{'}) \tag{2.10}$$

Where $W^{'}$, W, b, and $b^{'}$ are the parameters learned during the first two training phases. The output summary will be constituted of a set of $R_{summ} = \{x_1^{'}, x_2^{'}, ..., x_n^{'}\}$, where n is the average length of all reviews included in the batch.

2.4 Experiment

2.4.1 Dataset

We trained our model on the Amazon Product dataset composed of reviews of 82.83 million reviews for 9.35 million various items (He and McAuley, 2016). The products are divided into 29 distinct categories and include different metadata about users, product features, images, ratings, etc[2]. We select a subsample of reviews from fewer categories to avoid too large a vocabulary and respect our limited physical hardware capabilities. This choice is also motivated by the choice of our background autoencoder model for summarization. By generating an average representation of the input reviews, this model suffers when the semantic diversity becomes too important, producing broad and almost irrelevant summaries (Chu and Liu, 2019; Coavoux et al., 2019). Therefore, we sample around 18,000 opinions from mainly 2 categories: Cell Phone and Kindle Store Products, representing 80% of the reviews, and 3 additional categories: Pet supplies, software, and beauty items, accounting for the last 20% of the reviews' distribution. We continue filtering the dataset by not considering products having fewer than 15 reviews. We process the data to clean the text to reduce the number of end variables. Specifically, we removed emojis and URLs, we cleaned and unified encoding and special characters and whitespaces, and finally expands and normalize different contractions, expressions, or abbreviations. After this cleaning, we have used the Spacy tokenizer[3] and a review jas an average length of

[2] https://cseweb.ucsd.edu/~jmcauley/datasets.html

[3] https://spacy.io/api/tokenizer

60 tokens. We also remove products above the 90^{th} percentile in terms of the number of reviews to reduce bias from terms associated with these items. We ignore texts under 8 and above 200 tokens for the same reasons. Regarding the sentiment information, the reviews come with different ratings, indicating user satisfaction toward the product. Our dataset is composed of 22% rated 1, 19% rated 2, 19% rated 3, 20% rated 4, and 20% rated 5. We are aware of the limitations of considering these ratings as sentiment information but it is often representing real business cases, such as the net promoter score for the finance sector (Reichheld, 2006), and is ultimately convenient enough for becoming a common practice in the sentiment analysis literature (Fang and Zhan, 2015; Khoo and Johnkhan, 2018). Therefore we also transform the Amazon rating into three classes: negative sentiment for reviews rated 1 and 2, neutral for the ones rated 3, and positive for 4 and 5. We have paid particular attention to preserve a balance between positive and negative reviews and to filter neutral ones in the training and validation sets because the GRL could prevent the model from learning representations agnostic to neutral sentiment. Our final training, validation, and test splits consist of 11,096, 3,020, and 3,005 reviews. We will give more details about the annotation of the test set in the following section 2.4.3.

2.4.2 Implementation

Our model uses the GloVe 100 dimensional pre-trained word embeddings (version: glove.6B.100d) (Pennington et al., 2014). Both the encoder and the decoder of the model are composed of a single bidirectional layer with a size of 512 hidden units. We initialize the weight of the different layers via a Xavier uniform distribution (Glorot and Bengio, 2010), and we established the dropout of each layer at 0.2. We train the model for 250 epochs with the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-4} for the language model. We then fine-tune the model for 50 more epoch with the same optimizer and a new learning rate of 10^{-5} for the cosine similarity loss. For both cases, we use a weight decay of $8 * 10^{-3}$ and a gradient clipping of 10. We train and fine-tune the model with stochastic gradient descent using a gradient accumulation technique to obtain a batch size of 128. We apply the beam search method with a beam size set to five and an n-gram blocking method (Paulus et al., 2017) set to avoid trigram repetitions. We implemented our model with the Pytorch library[4] version 1.8.1. The model was trained on a machine with an 8 Gb NVIDIA Tesla P4 graphics card and a 60 Gb 16-core processor. Both the dataset and our code implementation are available on Github [5].

2.4.3 Evaluation metrics

To produce our test set and to evaluate our model, we used Amazon Mechanical Turk to develop a new dataset composed of 200 human-generated objective summaries, representing a total of 3,000 reviews. We asked the workers to create summaries that reflect the general opinion in a neutral/objective way. More specifically, the summary should be written as consensual review of all customers opinion and that, without omitting feelings, should be unbiased, and without personal judgment or point of view. Our dataset focus on our two main categories: cell phone and kindle items. For each product we sampled 15 reviews and we control the overall sentiment distribution to have 90 positive, 90 negative, and 20 neutral sets of reviews. We also asked the workers to produce summaries ranging from 20 to 80 words. To maintain the texts' quality, we employed the amazon filters to only have workers that have an approval rate above 85% and that have a basic language fluency in English. We also followed a two-step process to create the dataset. We first collected 5 summaries per product. We only approved summaries that respected the instructions in terms of both content and language quality. Then, using the same filter criteria on AMT, we requested a second panel of workers to identify the 3 most coherent texts. Although we are aware that this procedure does not guarantee that texts will be written exclusively by native speakers, this double-checking strategy has let us maintain good coherency and quality for our references.

As it is customary for document summarization evaluation, we report ROUGE-1, ROUGE-2, and ROUGE-L F1 scores (Lin, 2004) between our newly produced references

[4] https://pytorch.org/

[5] https://github.com/cd209392/neutral_summ

and the various systems evaluated. ROUGE-N is a metric that measures the overlap of n-grams between a model and reference summaries. ROUGE-L stands for Longest Common Subsequence and provides information about the longest co-occurring in sequence n-grams between the two texts. As we have 3 human summaries per batch, we disclose the average score between the systems and all 3 references, and the maximum score corresponding for its the best matching reference. As we attempt to analyze unsupervised methods, we also implemented different metrics quantifying objectivity and neutrality independently of human references. SentiWordNet (Baccianella et al., 2010) is a dictionary adopting a 3-dimension vector $[pos_{score}, neg_{score}, neu_{score}]$ respectively for the positive, negative, and neutral valence of a term. Authors in (Abdi et al., 2019a) have designed two metrics to assess news sentence subjectivity to produce factual summaries based on that dictionary. More specifically, they introduced a neutrality score by estimating the proportion of content words in a sentence where the main dimension is neutral, or not positive neither negative. They have also proposed an objective score for words: $ObjScore = 1 - (Pos_Score + Neg_Score)$, and once again determined the proportion of a sentence's objectivity. Therefore, we adopt their approach to measure the ratio of content words that are neutral and objective our generated summaries. To ensure the robustness of our subjectivity analysis, we have completed it with the subjectivity clues dataset submitted in (Wilson et al., 2009). It is a dictionary that collects terms with their associated subjectivity degree (weak or strong). This method is widely recognized for assessing document partiality and has also been applied for customer opinions to demonstrate, for example, their helpfulness (Singh et al., 2017). Once again, we report the proportion of these terms contains in models' output. Finally, to further improve the resilience of our evaluation, this time for neutrality, we have also examined the summaries' sentiment scores with VADER (Hutto and Gilbert, 2014). VADER (Valence Aware Dictionary for sEntiment Reasoning) is rule-based system that maps lexical features to emotion intensities and that is specialized for social media and web content. We report the number of summaries where the main compound was predicted as (neutral, positive, negative). While being quite simplistic, most of these metrics have demonstrated their capacity to assess the performance of models independently of any training in different contexts (Abdi et al., 2019a,b). Therefore, we believe that their combination provides a strong and robust way to analyze the neutrality and objectivity of automatically generated texts in our article.

2.4.4 Baselines

We compare our model with 3 different baselines:

- BERT for Text Summarization (Miller, 2019): This model is considered as a strong reference for extractive summarization. The BERT model is a pre-trained language model fine-tuned for summarization purpose. In this study we do not fine-tune further the model on our dataset to come close to an unsupervised configuration.
- TextRank (Mihalcea and Tarau, 2004): This unsupervised model is a graph based method for extractive summarization. The method demonstrated throughout the years it strong performances on multiple datasets and set-ups. This is also the model most commonly used as a baseline in the Text summarization literature.
- MeanSum (Chu and Liu, 2019): The model is an unsupervised abstractive model relying on a sequence to sequence architecture. Since it is the basis of our approach, we consider this model as a vanilla version that ignores neutral or objective information.

2.5 Results analysis

2.5.1 Model evaluation

The ROUGE results are presented in the table 3.1, and they are computed on our final test dataset described in section 2.4.3. Since we preserved 3 summaries per group of reviews, we report both the average score between the summary and all the references, and the maximum score with its best matching reference. The table shows that our model obtains comparable ROUGE results with the other abstractive approach. It indicates that the model can create a coherent and qualitative text despite the sentiment neutralizer bias.

Table 2.1: ROUGE scores on the Amazon objective dataset. R-1a, R-2a, R-La stand for averaged results between the summary and the 3 human references while R-1m, R-2m, R-Lm stand for the maximum results between the summary and the references.

Methods	R-1a	R-1m	R-2a	R-2m	R-La	R-Lm
BERT Summarizer (Miller, 2019)	**21,5**	25.6	2.1	4.2	11.6	15.2
TextRank (Mihalcea and Tarau, 2004)	19.6	**27.8**	**3.3**	**5.7**	14.5	17.9
MeanSum (Chu and Liu, 2019)	20.2	25.5	2.3	4.8	14.2	17.8
Our apprach	20.4	25.7	2.4	4.7	**15.5**	**19.4**

However, both methods are less performant than the extractive ones. It differs from the findings observed by the authors in (Chu and Liu, 2019), where the authors compare their models to the same extractive baselines and obtain better ROUGE F1 scores on their evaluation dataset. When we conpare the two evaluation dataset, we can formulate the following hypothesis. In our case, human assessors, when asked to come with objective and neutral review summaries that represent the overall opinion, failed to produce this review representing the input. Instead, they generate more a statement or a portrait of the main information. The human generated summaries provided in the table 2.3 emphasize this disposition. For example, sentences summaries such as “The story was short. It was a romance western. It was great for a beginner reader.” are essentially descriptive and do not act as a review. It thus favors extractive models that tend to select more descriptive and representative sentences, whereas unsupervised abstractive methods behave more like copycats of input reviews (Bražinskas et al., 2019), thus producing more affirmative sentences.

Table 2.2 reports the metrics for neutrality and subjectivity evaluation of the various approaches. We further exhibit the average score obtained by the 3 human references for these metrics. Results show that our model outperforms all the various baselines by generating more neutral and objective words and by getting closer to the human references as well. More specifically, SentiWordNet based metrics demonstrate that our method uses fewer words with a positive or negative dominant dimension. These results are corroborated by the VADER predictions where our model tends to employ more neutral terms than for the other baseline. Finally, it also produces fewer words with a subjective aspect,

Table 2.2: Neutrality / Objectivity evaluations on Amazon dataset.

Methods	Neutrality	Objectivity	% weak subjective	% strong subjective	VADER (neu,pos,neg)
Human references	94.2	89.4	2.3	1.92	(175,24,1)
BERT Summarizer	84.3	82.6	4.2	3.6	(184,23,0)
TextRank	85.9	87.8	5.9	4.9	**(201,6,0)**
MeanSum	84.2	87.1	4.8	2.9	(126,79,2)
Our apprach	**92.4**	**92.3**	**1.42**	**1.32**	(152,48,0)

either strong or weak. Our hypothesis is that, even if subjectivity is not directly linked to sentiment, the fact that we control its distribution when the training phase makes it possible to avoid subjective terms that are often associated with positive or negative sentiment. As one review is obviously a subjective representation, the extractive or abstractive methods, by depicting the main sentiment, will enforce that potential subjectivity at least in terms used. Overall, these evaluations demonstrate that our model is closer to the human behavior when it comes to creating a summary that does not omit feelings but provides a constructive view without personal judgment.

Finally, some examples of generated summaries by our model, the baselines, and the references are provided in the table 2.3. In these examples, we can observe the differences between the various outputs. First, we can note the descriptive aspects of the extractive strategies with sentences like “Love this book” or "No depth to story." rather than using personal words like “I” that fit the same kind of expression like “the book was nice” in the references. If we look at the different summaries, we can also see that our approach favors the use of neutral informative terms such as "characters" or "read", closing the gap with the extrative summaries in that sense. We can see that the MeanSum abstractive baseline succeeds in its main purpose, which is to convey the user’s feelings about their experience. On the contrary our approach favors words like “interesting” or "good", which still carry sentiments, but are much less intense or subjective than terms as “best” or “great” as observed for the abstractive model or "Love", "like", or "nice" in the extractive summaries.

We conclude that our model achieve the goal of producing summaries that are neutral

Table 2.3: Example of generated texts. Table with examples of generated text for 2 products. For each product we provide the 3 human references with the text generated by our model, the absractive model Meansum and the extractive model TextRank.

B00K7Q8I1A	Human references	A short book about love. it ends at a cliffhanger meaning there will be another book in the future. **The book was** nice to read but one have buy next book to see the ending. This erotic book is good overall. However, the story is short, which leads to a rushed ending that can be disappointing. The end of the story leaves the reader waiting for a second episode.
	Our model	the book was **interesting** and the ending was interesting i was left hanging on the book and i could not wait to read more
	MeanSum	i was not sure if it was the norm. i am not sure if i can find this album.
	TextRank	Soon as they are discovered she takes off and the book ends.... **Love this book** and waiting to read the other two. Sadly, even the little teaser at the end isn't enough to salvage what could've been a good story.
B00ARZIOE2	Human references	The story was short. It was a romance western. It was great for a beginner reader. A very short story of love with good passion but it needed some depth in characters and the story to be a novel. Good read for short story lovers but it's not for novel readers. it's a short dramatic story for a quick read. good for reading on a plane or taxi
	Our model	i really enjoyed this book it was written a good read and i did not like the characters and the characters were so it was not a good read
	MeanSum	i was expecting a good book. it was not the **best**, but it did not fit my needs. it was a little difficult to get a seat and it was not that **great**
	TextRank	Of the hundreds of books I have read over the past few years, this is the most poorly written one of them all. This book is extremely short and while a clever story line, hardly time to get interested before this mini novel ends. The book was very good I like this type of romance stories.

and objective while preserving relevant content for analyzing customer reviews. Since we have used Meansum as a base to study the impact of information bias in text generation, the objective is not to resolve the core problem of this baseline regarding coherence or hallucinations. However, we still note a slight degradation with the increase of hallucination terms that we detail in the limitation section 2.5.3 of the article.

2.5.2 Model and configuration analysis

For the purpose of better understanding our method, we offer a description and an analysis of various possible configurations during the training and the generation of the summaries. As a matter of fact, our approach includes three steps where different representations choices can be made and that can impact the final solution.

- During the adversarial review reconstruction training phase, we can either decide to train the model independently from the sentiment information by using the original review encoding H. We can also employ the sentiment-specific $\tilde{H}$, or a concatenated representation of $\hat{H}$ and $\tilde{H}$ that we named H_{concat}.
- Throughout the fine-tuning steps, we can decide to fine-tune the model towards sentiment agnostic representation by reducing the similarity of the summary generated by H or $\tilde{H}$ with the review expressed respectively by $\hat{H}$ or H. We can also inverse the fine-tuning process towards sentiment specific representation. Finally, we can choose to not bias the model by simply considering the similarity between the same representations.
- depending on the options made in the training and fine-tuning steps, it is thus interesting to study the production of the summaries once again with H, $\tilde{H}$, or $\hat{H}$ (or possibly H_{concat}) either used for the hidden or the context representation in equation 2.2.

All these arrangements have led to various outcomes. Since there are potentially around 200 different existing configurations and that some of them did not induce exploitable outputs, we present the comparison of the best and most compelling results. The following list provides, for each configuration analyzed, the training layers, the fine-tuning layers for producing the summary, the layers for encoding the reviews, and the representation used for the summary generation.

- Configuration 0: This is our current configuration used for the previous results. Training: $\tilde{H}$; fine-tuning: $H \rightarrow \hat{H}$; summary generation: H.
- Configuration 1: Training: $\tilde{H}$; fine-tuning: $\hat{H} \rightarrow H$, summary generation: H.
- Configuration 2: Training: $\tilde{H}$; fine-tuning: $\hat{H} \rightarrow \tilde{H}$, summary generation: H.
- Configuration 3: Training: $\tilde{H}$; fine-tuning: $H \rightarrow H$, summary generation: H.
- Configuration 4: Training: $\tilde{H}$; fine-tuning: $\tilde{H} \rightarrow \tilde{H}$, summary generation: H.
- Configuration 5: Training: H; fine-tuning: $H \rightarrow H$, summary generation: H.
- Configuration 6: Training: H; fine-tuning: $H \rightarrow \hat{H}$, summary generation: H.
- Configuration 7: Training: H; fine-tuning: $\hat{H} \rightarrow H$, summary generation: H.
- Configuration 8: Training: H_{concat}; fine-tuning: $H_{concat} \rightarrow H_{concat}$, summary generation: $\hat{H}$.
- Configuration 9: Training: H_{concat}; fine-tuning: $H_{concat} \rightarrow H_{concat}$, summary generation: H.
- Configuration 10: Training: H_{concat}; fine-tuning: $H_{concat} \rightarrow H$, summary generation: H.

The table 2.4 includes results for the average ROUGE F1 score with the human references as well as the various metrics for neutrality and objectivity for the comprehension of the model's behavior.

The analysis of the different procedures allows us to draw several interesting conclusions on how the model captures information and sentiment aspect in reviews. First, it is important to mention that we have tried a configuration where we train the language model with $\hat{H}$ alone, but it did not have a sufficient capacity to recreate reviews. Our hypothesis is that if we directly exploit the representation from the gradient reversal layer $\hat{H}$, the model enters a competitive mode between the objective of classification and reconstruction.

Table 2.4: Configuration analyses. Table introducing the different results from various model configurations. We repeat the results of our main model in the first line for comparison. Bold results indicates main aspects to account for each other configuration.

Configuration	R-1 (F1)	Neutrality	Objectivity	% weak subj .	% strong subj.	VADER
Configuration 0	20.4	92.4	92.3	1.4	1.3	(152,48,0)
Configuration 1	20.6	77.0	82.1	3.8	1.8	(107,92,1)
Configuration 2	17.0	66.9	66.7	4.9	0.2	(5,195,0)
Configuration 3	18.7	73.3	82.1	5.05	2.41	(51,149,0)
Configuration 4	20.3	76	80.6	3.76	1.56	(96,110,30)
Configuration 5	20.1	93.7	89.3	2.65	1.7	(155,44,1)
Configuration 6	20.0	92.7	89.0	2.8	1.8	(152,47,1)
Configuration 7	17.4	98.4	99.2	0.4	0.7	(199,1,0)
Configuration 8	13.1	96.1	65.3	4.5	12.2	(196,4,0)
Configuration 9	15.5	92.2	89.7	6.2	4.8	(181,19,0)
Configuration 10	14.5	94.6	95.2	4.4	12.14	(207,0,0)

Without the decorrelation of the layers with the projection mechanism, this makes learning the language model practically impossible. By using a concatenation of both the sentiment representations $\tilde{H}$ with $\hat{H}$ in H_{concat}, the model then gets signal enough to perform on the two objectives. This approach at least allow us to analyze the impact of $\hat{H}$ on the learning process.

Regarding the results, we observe the importance of the projection mechanism in the initial training step, as defined in equation 2.6. We notice that if we train the reconstruction with $\tilde{H}$, the model encapsulates sentiment material in the correct layers. In these configurations, we can assess the impact of the fine-tuning strategy on information bias. Specifically, the fine-tuning tested for configurations 2 and 3 show that if we generate the summary with $\hat{H}$, our sentiment agnostic representation, and constrain it towards a representation that conveys more sentiments, we decrease the neutrality and the objectivity of the texts, producing the opposite outcomes of our current approach, as we anticipated. The complementary observation of the configurations 0, 3, and 4 results also emphasize the role of the projection mechanism defined in 2.8. Indeed, if we do not use this projection mechanism for creating the representation $\hat{H}$, then the model is not able to reduce nor the subjectivity or the neutrality of the summaries. These results confirm that both projections are valuable in the learning process and that $\tilde{H}$ captures sentiment biased information while $\hat{H}$ seems to encode agnostic sentiment content. Moreover, if the fine-tuning is performed with the same layers, thus not influencing sentiment information during that step, as for configurations 4 and 5, the model is not able to bias the text generation, replicating our baselines such as MeanSum. The results from configurations 5,6, and 7, where the model has been trained directly with the reviews' encoding H, demonstrate that it is necessary to employ a sentiment-dependent representation in the reconstruction process so that the fine-tuning and the summary generation yields coherent and expected outcomes. Especially, when we fine-tune the model from the H towards the sentiment agnostic representations $\hat{H}$, we assume it would increase neutrality and objectivity of the output and conversely if we reverse the fine-tuning. However, we even observe opposite response in terms of neutrality and objectivity from what we could expect. It highlights a crucial characteristic feature of the joint adversarial procedure. It ensures to reduce the difference between domain hyperspaces, thus providing a better alignment of the projected representation $\tilde{H}$ with H in the sentiment space. We understand that employing the gradient reversal layer independently by training the language model with H, does not constrain these two hyperspaces conducting to create potentially misleading representations. We can also note that it is easier in this set up to bias information towards neutral outputs rather than sentiment-oriented ones, leading us again to think that hyperspaces are no longer aligned.

We introduce supplementary observations worth mentioning to understand better the functioning of this model. We can see that if we use the sentiment agnostic representation $\hat{H}$ in the learning of the language model with H_{concat} as for configurations 7, 8, and 9, we obtain incoherent summaries. It corroborates our hypothesis that the projection mechanism provides a necessary decorrelation between the two different objectives to let the language model to freely learn to reproduce reviews. Moreover, during these experiments, we even had to start the training of the sentiment classifier 5 to 10 epochs after training the language model, otherwise the influence will be too strong to stabilize the learning. Another interesting phenomenon to observe with these configurations is that with every fine-tuning procedures used for the generation we produce texts that achieve neutral scores, where VADER model is even unable for certain to predict anything else than neutral summaries, but with a high utilization of subjective terms. The reading of the texts shows a balanced output of extremely positive and negative terms. It leads us to make us think that $\hat{H}$ actually captures sentiment extremity in both positive and negative spectrum. Having a proper learning of the agnostic space thus mitigate the use of intense and subjective terms explaining why we succeed in obtaining not only neutral summary but also more objective ones in our best configuration. Finally, our last observation relates to the ROUGE score and the coherence of the produced outputs in general. We indeed denote a performance decrease when we do not apply H in the generation process whatever the configuration used. This emphasizes the fact that H apprehends the grammatical and coherency structure in texts whereas the other representations such as $\tilde{H}$ or $\hat{H}$ predominantly capture and bias information meaning. This once again is possible because these representations are separated through the projection of layers and not combined directly together.

Finally, to understand how pre-trained generative models perform on the task and how our approach may be complementary, we asked ChatGPT to produce objective summaries on our two examples. We provided the same instructions to the model as those given to AMT workers, i.e., "generate summaries that reflect the consensus opinion in the reviews and that, without omitting feelings, should be unbiased, and without personal judgment or points of view with a length equal to the average length of input reviews". We can observe that the model always starts with the same non-informative general sentence about the consensus opinion of mixed reviews. Moreover, the model tends to report all the facts to preserve balance in the text instead of depicting the representative opinion objectively and thus creates repetition and paradoxical information. It is unable to reduce further the size of the summary despite the prompt requesting to do so. This fundamental difference can be easily noticed by comparing empirically with human summaries. Indeed, we observe the same balance of positive and negative elements but with fewer repetitions of the same aspects to reduce the size of the summary. Obviously, We know the limitations of this test we have performed here, especially when no real prompt engineering has been done, especially without testing the number of prompts to obtain the best results. These two simple examples are included to introduce potential challenges faced by such models. Moreover, these observations and the results from the BERT summarizer highlight how adversarial methods and information bias in intermediate representations layers such as proposed in our approach could contribute to improving pre-training or fine-tuning general large language models to better identify the information and to answer better to specific tasks. We let the empirical study and comparison of the performances of such generative approaches and their associated prompting techniques to future research.

2.5.3 Model limitations

The comparison with the different baselines and configurations have also highlighted some limitations of our approach. The sentiment constraint implied by $\tilde{H}$ when learning to reconstruct the reviews in the training steps, biases the model to employ the most occurring neutral words of the input. We make this observation with terms like "story", "book", or "case" that are overused in the summaries even for the wrong product category. We first observe that these terms belong to the most occurring ones along with grammatical words but also words as "product" or "phone". The major difference we note between them is their sentiment distribution. While "product" and "phone" appear approximately 45% of the time in both negative and positive reviews and 10% in neutral reviews, the 3 other words "story", "book", or "case" occurred in neutral reviews around 20% to 25% of the time. We think that the model considers those words are both relevant and sentiment agnostics, being overconfident when it comes to employ them, and thus polluting the outputs and increasing hallucinations compared to our baseline model.

This also confirms our second limitation, where we need need to perfectly control the sentiment distribution in the training sets. The presence of neutral words in both positive and negative reviews, or the different projection mechanisms impact the learning process of the model and how it affects the quality and the information bias in the generated summaries. Moreover, sentiment distribution is not directly correlated to objectivity, and sometimes create dissonance between the sentiment reduction and the use of subjective

Table 2.5: Objective summarization with ChatGPT. Examples of objective summaries generated by ChatGPT when provided with the same instructions as the Amazon workers. The same 2 books products have been used for the tests.

Book 1	ChatGPT	The consensus opinion in the reviews reflects mixed reactions to the book. Some readers found the story unimpressive, lacking depth and suspense, and felt it was predictable and not worth their time or money. They expressed a desire for a stronger story line and more engaging elements. Others criticized the poor writing quality and short length of the book, feeling it didn't allow enough time to become invested in the story or characters. On the other hand, positive reviews highlighted the appeal of the story as a Western romance and praised its believability and enjoyable nature. While some readers desired more depth in character development and plot, they still found it to be a well-written and engaging love story. The short length of the book was appreciated by some, making it a quick and convenient read for travel or spare moments. Overall, opinions varied, with criticisms focusing on the lack of depth and length, while positive reviews emphasized the genre's appeal and the book's entertainment value.
Book 2	ChatGPT	The consensus opinion in the reviews reflects a mixed response to the book. Some readers enjoyed the story but felt that it could have had a stronger and more developed storyline. The unresolved ending left others feeling dissatisfied and eager for a follow-up.There were criticisms regarding the brevity of the novella, with concerns about the rushed plot and stunted character development. Several reviewers expressed disappointment with the unexpected and unfinished ending, considering it to be too short and leaving them wanting more. However, positive reviews praised the book as amazing, entertaining, and enjoyable. Some readers expressed excitement and anticipation for the next books in the series. It's worth noting that the cliffhanger ending received mixed reactions, with some disliking it and others finding it enticing for future installments. Overall, opinions varied, highlighting the book's strengths and weaknesses, leaving readers with different levels of satisfaction and anticipation for what comes next

words, especially those with strong connotations.

Another limitations comes from the model we have used. By introducing the adversarial objective, we incorporate many additional hyperparameters and new layers to play with on our model. We found that the performance of the model and the observations we have made above are quite sensitive to these hyperparameters. This raises issues of stability and portability of our model to apply it to new contexts.

Finally, we are aware that our model employs a basic approach with a single-layered autoencoder and a GRL for disentanglement. This choice was made partly made by our hardware limitations but also motivated by the need to easily interpret the results. One issue with employing large language models for studying information bias is that we do not know how the information has been encapsulated in such complex architecture, especially during pre-training steps. Therefore, it would have greatly limited our ability to analyze how to enforce text objectivity. We also recognize that we do not apply the most recent or advanced baselines to compare our results, and this decision is driven by the same reason. By focusing on simpler models that have been widely used in unsupervised text summarization, it allows a better comprehension of how classic models behave for this new task and thus let other researchers to easily judge our approach and interpret our results.

2.6 Conclusion to chapter 2

Within the framework of this study, we hope to have illustrated the importance of producing objective opinion summaries. We introduce a novel unsupervised automatic document summarization method based on gradient reversal layers. Moreover, we proposed a new dataset as well as new metrics to analyze the summaries. We have demonstrated the interest of adversarial learning to bias information in generative text models. In future work, we plan to study how to apply this approach to large multi-layer transformer architectures. We are also enthusiastic in exploring how to create representations that are either contrastive or agnostic other variables such as genders, languages, or product domains. Meanwhile, we believe that we have proposed an interesting baseline for this original summarization task that opens the door to new research avenues that are still unattended in the field.

References

Abdi, A., Shamsuddin, S. M., Hasan, S., and Piran, J. (2019a). Automatic sentiment-oriented summarization of multi-documents using soft computing. *Soft Computing*, 23(20):10551–10568.

Abdi, A., Shamsuddin, S. M., Hasan, S., and Piran, J. (2019b). Deep learning-based sentiment classification of evaluative text based on multi-feature fusion. *Information Processing & Management*, 56(4):1245–1259.

Ahmet, A. and Abdullah, T. (2020). Recent trends and advances in deep learning-based sentiment analysis. *Deep learning-based approaches for sentiment analysis*, pages 33–56.

Amplayo, R. K., Angelidis, S., and Lapata, M. (2021). Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Angelidis, S. and Lapata, M. (2018a). Multiple instance learning networks for fine-grained sentiment analysis. *Transactions of the Association for Computational Linguistics*, 6:17–31.

Angelidis, S. and Lapata, M. (2018b). Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the*

Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA).

Baron, R. A. (1993). Criticism (informal negative feedback) as a source of perceived unfairness in organizations: Effects, mechanisms, and countermeasures.

Blitzer, J., Dredze, M., and Pereira, F. (2007a). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

Blitzer, J., Dredze, M., and Pereira, F. (2007b). Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

Bražinskas, A., Lapata, M., and Titov, I. (2019). Unsupervised opinion summarization as copycat-review generation. *arXiv preprint arXiv:1911.02247*.

Cao, Z., Li, W., Li, S., and Wei, F. (2017). Improving multi-document summarization via text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Chaturvedi, I., Cambria, E., Welsch, R. E., and Herrera, F. (2018a). Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion*, 44:65–77.

Chaturvedi, I., Ragusa, E., Gastaldo, P., Zunino, R., and Cambria, E. (2018b). Bayesian network based extreme learning machine for subjectivity detection. *Journal of The Franklin Institute*, 355(4):1780–1797.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8,* *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Chu, E. and Liu, P. (2019). Meansum: a neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*, pages 1223–1232. PMLR.

Coavoux, M., Elsahar, H., and Gallé, M. (2019). Unsupervised aspect-based multi-document abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Colhon, M., Vlăduţescu, Ş., and Negrea, X. (2017). How objective a neutral word is? a neutrosophic approach for the objectivity degrees of neutral words. *Symmetry*, 9(11):280.

Fang, X. and Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, 2(1):1–14.

Ganesan, K., Zhai, C., and Han, J. (2010). Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 340–348, Beijing, China. Coling 2010 Organizing Committee.

Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Gui, L., Jia, L., Zhou, J., Xu, R., and He, Y. (2020). Multi-task learning with mutual learning for joint sentiment classification and topic detection. *IEEE Transactions on Knowledge and Data Engineering*.

Havaei, M., Mao, X., Wang, Y., and Lao, Q. (2021). Conditional generation of medical images via disentangled adversarial inference. *Medical Image Analysis*, 72:102106.

He, R. and McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Jang, E., Gu, S., and Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Khoo, C. S. and Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4):491–511.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipfelsberger, P., Herhausen, D., and Bruch, H. (2016). How and when customer feedback influences organizational health. *Journal of Managerial Psychology*, 31(2):624–640.

Krishnan, J., Purohit, H., and Rangwala, H. (2020). Unsupervised and interpretable domain adaptation to rapidly filter tweets for emergency services. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 409–416. IEEE.

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

Liu, J., Cao, Y., Lin, C.-Y., Huang, Y., and Zhou, M. (2007). Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 334–342, Prague, Czech Republic. Association for Computational Linguistics.

Lovinger, J., Valova, I., and Clough, C. (2019). Gist: General integrated summarization of text and reviews. *Soft Computing*, 23:1589–1601.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

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

Miller, D. (2019). Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Paulus, R., Xiong, C., and Socher, R. (2017). A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Pecar, S. (2018). Towards opinion summarization of customer reviews. In *Proceedings of ACL 2018, Student Research Workshop*, pages 1–8, Melbourne, Australia. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Reichheld, F. (2006). The ultimate question: Driving good profits and true growth. *Boston, MA*.

Seng, D. and Wu, X. (2023). Enhancing the generalization for text classification through fusion of backward features. *Sensors*, 23(3):1287.

Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y. K., Saumya, S., and Roy, P. K. (2017). Predicting the “helpfulness” of online consumer reviews. *Journal of Business Research*, 70:346–355.

Suhara, Y., Wang, X., Angelidis, S., and Tan, W.-C. (2020). OpinionDigest: A simple framework for opinion summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5789–5798, Online. Association for Computational Linguistics.

Tang, H., Mi, Y., Xue, F., and Cao, Y. (2021). Graph domain adversarial transfer network for cross-domain sentiment classification. *IEEE Access*, 9:33051–33060.

Tsytsarau, M. and Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24:478–514.

Whetten, D. A. and Cameron, K. S. (2005). Developing management skills. *(No Title)*.

Wilson, T., Wiebe, J., and Hoffmann, P. (2009). Articles: Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.

Chapter 3

Topically diversified summarization of customer reviews

Long version of the article accepted at the *ICNLSP 2023* Conference

Abstract

Promoting information coverage while reducing redundancy has always appeared as a crucial issue for handling data heterogeneity in multi-document summarization. Diversity of topics addressed in summaries is an efficient method to tackle this challenge. We introduce a self-supervised algorithm for multi-document summarization that employs a multitask learning approach for topic diversification. Our model is based on two variational autoencoders that combine the training of a language model and a topic model. We then use the topic distribution to control content in the generated texts by the language model. We evaluate our method on the Amazon product review dataset, and we report ROUGE results and other metrics such as BLEURT scores to assess information coverage. We demonstrate that our approach not only creates different output for the same batch of reviews but also optimizes our evaluation metrics. Our study finally emphasizes how we can apply topic modeling to bias information in autoencoder models and how various

strategies let us produce either diversified or aspect-focused summaries.

3.1 Introduction

E-commerce and online sales platforms have grown substantially over the past years and have become the main shopping media in recent years. These platforms bring a major change in the way we purchase products or services, as they allow us to share our experience and access to that of other users. Instead of accounting for company descriptions that might be biased, consumers are relying more and more on others' recommandations and advice. However, to make an informed decision, due to their subjective nature, customers must read many reviews to get an idea of the product quality. Automatic text summarization is the process of distilling the most important content to create a reduced version of these opinions, therefore becoming crucial to help online platform users.

The recent success of deep learning systems has led to significant improvement of extractive (Angelidis et al., 2021) or abstractive (See et al., 2017; Paulus et al., 2017) document summarization models. In the framework of customer opinions, one of the major issues is their extreme subjective and domain sensitive nature. This makes the production of large parallel corpora costly and hardly transferable, which has always created a strong appetite for unsupervised summarization approaches. In this context, the definition of salient information becomes crucial to meet the users' needs. For opinions, relevant content is often considered as consensual or representative of the general point of view. In the abstractive framework, this consists in generating a new opinion based on the average representation of all the individual reviews (Chu and Liu, 2019; Bražinskas et al., 2019). However, this approach suffers greatly from the topics and aspects diversity as well as from the potential contradictory non-factual information present in the corpus, hence producing an overly broad opinion (Coavoux et al., 2019; Amplayo et al., 2021). It is therefore essential to design aspect-based summarizers since they allow to capture different topics and entities to bring more fine-grained content into the summary. The strategy consists in creating groups of aspects through clustering techniques then extracting the relevant opinions related to them (Pecar, 2018). Abstractive methods take up this methodology by using clusters as input to pre-trained or fine-tuned language models to condition text generation (Suhara et al., 2020).

Considering the structure of a single review, we notice that many aspects can be described implicitly and which can be grouped together according to broader themes. The task of aspect retrieval has therefore been naturally associated with the detection of topics or subtopics of a product review (Zhai et al., 2015). Therefore, exploring aspects through topic models presents the advantage to infer these groups of them dynamically without any labeling necessary. Methods such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), and its deep learning variants, have proven to be particularly efficient in identifying these themes for opinion datasets (Ozyurt and Akcayol, 2021; Xiao et al., 2018). At the corpus level, the different needs of the user's population result in the requirement to cover as many topics as possible not to miss any important aspect. In the context of opinion summarization, it is then essential to balance between the coverage of the semantic space and the diversity of the topics to incorporate the largest volume of information in the final summary. Most extractive approaches employ optimization framework for selecting sentences that address the main themes while penalizing their intrinsic similarity (Yogatama et al., 2015; Li et al., 2010; Fang et al., 2015). Early abstractive summarization methods showed that weighting the attention given to terms according to the similarity, thus enforcing the diversity of the learned representations, allowed models to cover better the content material of a heterogeneous input (Fabbri et al., 2019). Finally, in an unsupervised setting, Conditional Variational AutoEncoders (CVAE) (Sohn et al., 2015) combined with generative topic modeling systems have demonstrated interesting capacities for producing sentences with varied topics (Gao and Ren, 2019; Xiao et al., 2018).

In this article, we introduce a method for unsupervised customer opinion summarization that can either generate a text focused on a particular topic/aspect or that maximize input coverage. More specifically, our model relies on a multi-objective function approach to train jointly a language and a topic model. The topic model is trained with a first variational autoencoder (VAE) establishing a Dirichlet distribution from a general bag of words model. The distribution devises a latent representation that conditions the review reconstruction in the second VAE. During the generation phase, we can select a subset of topics to first mask information input then to bias the final summary depending on the user needs. We performed the evaluation of our approach on the Amazon product reviews dataset (Bražinskas et al., 2019), demonstrating the importance of topic modeling to bring detailed and meaningful content in such a heterogeneous context.

3.2 Related work

3.2.1 Multidocument Summarization for Opinion

For opinions, multiple authors consider that the summary should be a consensual depiction of the input reviews and therefore that salient information must express popular content. The extractive methods of TextRank (Mihalcea and Tarau, 2004) relies on a similarity graph in order to elect the most representative sentences. Opinosis (Ganesan et al., 2010) runs through a co-occurrence graph to find the shortest and most redundant path to create semi-abstractive outputs. Meansum (Chu and Liu, 2019) is an unsupervised abstractive model where the summary results from the average of each review latent representations. Finally, authors in (Bražinskas et al., 2019) push the idea by exploiting a hierarchical generative VAE to produce the summary. However, such models suffer from the major shortcoming of causing overly broad summaries that become almost irrelevant if a lot of aspects are addressed in the opinions. The objective is then to develop a method that founds aspect-based information and ensure that the summary cover as many as possible. In (Angelidis and Lapata, 2018), authors create aspects representations with a partial autoencoder and devise an optimization function to select opinion that maximizes their coverage. As aspects are related to topics, some approaches suggested to explicitly extracts label through classification (Isonuma et al., 2017) or clustering such as OpinionDigest (Suhara et al., 2020), and then design once again ranking models to maximize their presence in the output. Finally, (Amplayo et al., 2021) introduces an interesting method that produces both general and aspect-specific summaries by clustering opinions and extracting sentences based either on popular or particular aspects. Regarding abstractive summarization, authors in (Coavoux et al., 2019) combines Meansum (Chu and Liu, 2019) with a clustering algorithm to devise latent representation and create different text for each group to maximize input coverage. Our model is closely related in the way that we modify the hierarchical VAE submitted in (Bražinskas et al., 2019) with a generative topic model. Our approach diverges because we propose a multi-task learning objective to enhance the dynamic detection of topic when producing summaries. Finally, this method also enables us to condition text generation based on popular aspects or a particular act of improving the model capability to satisfy the needs of all users.

3.2.2 Topic modeling

Topic modeling has been applied to all types of datasets or tasks to identify the prevalent themes of a corpus. Automatic text summarization is no exception; the first model extracts the sentences associated with the main topic, which results from a singular value decomposition (Gong and Liu, 2001). However, the rise of topic modeling begins with the development of the Latent Dirichlet Association (LDA) (Blei et al., 2003), a generative probabilistic model, and we refer to this article for a detailed depiction of historical topic modeling approaches. The authors in (Arora and Ravindran, 2008) submit an extractive summarization method by using these two properties to propose either inferential or generative models to select the most plausible sentences affiliated to the topics that also have the highest probability of being present in the corpus. To increase the coverage of the input texts, the importance of the LDA topics can then be weighted by their similarity to ensure the diversity of extracted sentences (Ren and De Rijke, 2015). Recently, variational autoencoders (VAEs) (Kingma and Welling, 2013), have achieved encouraging results in various NLP areas thanks to robust latent encoding of the entire sequence and to produce coherent outputs (Bowman et al., 2015). The training of the variational autoencoder relies on the re-parameterization trick gradient backpropagation with a prior Gaussian distribution.

Thereafter, different methods have adapted this trick to multinomial distributions such as the Dirichlet distribution, thus allowing the emergence of generative topic models such as AVITM (Srivastava and Sutton, 2017). With this approach, for a set of documents, it is possible to produce topically biased latent representations by weighting input information by topics (Gao and Ren, 2019), or to concatenate the latent representation directly with topic vectors (Xiao et al., 2018) to obtain conditional language models, in both cases, to diversify outputs sentences for the same input. Our approach combines these two approaches to generate topic-biased representations for an ensemble of opinions. Our approach either creates multiple abstractive summaries, each focused on a specific topic, or selects several topics weighted by their similarities to increase the aspects covered in the summary.

3.3 Proposed Model

This section presents the general architecture of our approach. Following the system introduced in (Bražinskas et al., 2019), we use a hierarchical VAE to encode both individual and group reviews into a latent semantic space. Our first main contribution is to combine this model with a second VAE for modeling topics as depicted in (Xiao et al., 2018). The architecture of our model is presented in the figure 3.1. Specifically, the VAE trained with a bag of words representation encodes the topic distribution through the latent variable t. The second VAE encodes the group review through the latent variable c. The latent variable z encodes individual content and is also conditioned by c. It is then combined with the topic variable t to generate the original review. We train the model with a multi-task objective function to produce a topically condition language model. Our second contribution occurs during summary generation where we introduce methods to select the most important topics, to bias input information, and to condition text regarding the chosen topic.

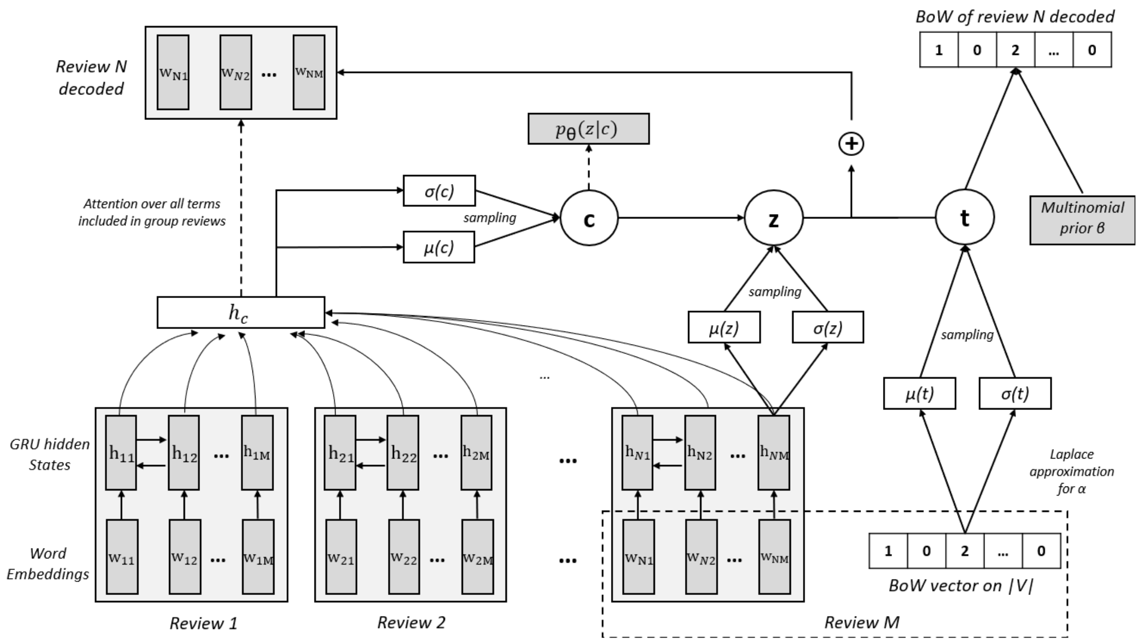

BoW of review N decoded
1 0 2 ... 0
Review N decoded
w_{N1} w_{N2} ... w_{NM}
$p_\theta(z|c)$
σ(c)
μ(c)
sampling
c
z
t
Multinomial prior β
Attention over all terms included in group reviews
h_c
sampling
μ(z)
σ(z)
sampling
μ(t)
σ(t)
Laplace approximation for α
GRU hidden States
Word Embeddings
h_{11} h_{12} ... h_{1M}
h_{21} h_{22} ... h_{2M}
h_{N1} h_{N2} ... h_{NM}
w_{11} w_{12} ... w_{1M}
w_{21} w_{22} ... w_{2M}
w_{N1} w_{N2} ... w_{NM}
1 0 2 ... 0
BoW vector on |V|
Review 1
Review 2
Review M

Figure 3.1: General model architecture. Multitask objective architecture for topic diversification of summary generation. The right part presents the VAE for topic modeling. The left part displays the second conditioned autoencoder that learns the language model.

3.3.1 Model background

The general architecture of our approach is based on two sub-models whose respective objectives are to learn a topic model and a language model. Each framework employs its own variational autoencoder to produce relevant latent representations. They are combined to condition text generation by on one or more topics. The corpus is composed by an ensemble of customer reviews on different products. The total vocabulary of the corpus is denoted V. Let's define a batch of M customer reviews regarding a specific product as $\{R_1,...,R_i,...,R_M\}$ used to train our model. Each review R_i is composed of a set of words $X = \{X_1,...,X_j,...,X_N\}$, where N represent the variable length of each review.

Topic Model

We start by applying the Bag of Words encoding method to represent each review. For a given review R_i, we obtain a vector BoW_i where each dimension indicates the frequency f_i of appearance of words in the review. To focus on product aspects, we employ Spacy

Part-of-Speech tagger [1] to process input texts and to only consider nouns. This vector is used as the input of a two-layer Feed Forward Neural Network with a softplus activation function. We then get a dense representation h_i^{bow}. We use h_i^{bow} to generate the continuous latent representation t_i observing a Dirichlet distribution and thus encoding the topics addressed in each review. Finally, our goal here is to maximize:

$$\log \int \prod_{i=1}^{M} p_\theta(BoW_i | t_i, \beta) \tag{3.1}$$

Where β is the multinomial prior distribution matrix of the topics over the vocabulary. Subsequently, we follow the method used for the ProdLDA model (Srivastava and Sutton, 2017), where we can approximate the mixture of two multinomial distributions to their weighted multiplication. Therefore, we can then multiply the β matrix with our topic vector t_i to compute the probability of generating the output Bag of Words $BoW_i^{'}$:

$$p_\theta(BoW_i^{'}) = softmax([t_i \cdot \beta]) \tag{3.2}$$

We train this part of the model with the mean square error function.

Language Model

We transform each input review with an pretrained embedding model to obtain a dense representation $E_i == \{E_{i1}, ..., E_{ij}, ..., E_{iN}\} \in \mathbb{R}^{*\mathbb{N}}$, where d is the embedding space dimension. The embedding matrix is then fed to a bidirectional Gated Recurrent Unit (GRU) (Cho et al., 2014), producing a representation h_{ij} for each word $j \in R_i$. Then last hidden state output h_{iN} is also used as the sentence representation.

For learning the language model by performing review reconstruction, we reproduce and adapt the hierarchical structure of VAE proposed in *Lsumm* (Bražinskas et al., 2019).

[1] `https://spacy.io/usage/linguistic-features`

Therefore, we first create a hidden representation h_c for all the group. We then compute the attention given to the set of words composing the reviews:

$$a_{ij} = softmax(v^\top \tanh(W_c, m_{ij}, b_c)) \tag{3.3}$$

where $m_{ij} = [h_{ij}; E_{ij}]$ is the concatenation of the embedding and the GRU representations of the word x_j of the review R_i, and W_c and b_c are learnable parameters of the model. Then we compute the hidden representation as a weighted sum over the attention of each word:

$$h_c = \sum_{i=1}^{M} \sum_{j=1}^{Ni} a_{ij} \tag{3.4}$$

We then assume a normal Gaussian distribution and apply a linear projection on h_c to sample the latent representation c encoding the information from the review group. We seek here to reconstruct the review R_i, therefore employing the same procedure and distribution assumption, we use a concatenation of h_{iN}, the last GRU layer of R_i, and c to sample the latent variable z encoding individual material from the review.

When reconstructing, we perform N decoding steps to generate our sentence. We start by fixing the initial hidden state of the decoder s_0 to $[z_i; t_i]$ the concatenation of the topic and latent representation of R_i, and, at each decoding step t, a simple GRU decoder estimates the current hidden state s_t with the states s_{t-1} and the predicted word $x^{'}_{t-1}$ at preceding steps:

$$s_t = GRU(s_{t-1}, x^{'}_{t-1}) \tag{3.5}$$

At decoding step t, the attention distribution $a^t_{\_i}$ is calculated over every other review of the group $R_{\_i}$, excluding R_i, as in (Bahdanau et al., 2014):

$$a^t_{\_i} = softmax(v^\top \tanh(W_{\_i}h_{\_i}, W_s s_t, b_{attn})) \tag{3.6}$$

where $W_{\_i}$, W_s, and b_{attn} are learnable parameters of the model. Once the attention is calculated, each individual attention value $a^t_{\_i}$ is used to weight the representation $h_{\_i}$ of each word not belonging to R_i, allowing the model to create a contextual representation c_t that focus on specific words at every step:

$$c^t = \sum_i a^i_t h_{\_i} \tag{3.7}$$

The context vector is concatenated with the decoder state and passed through a linear and a softmax layer to compute the probability of generating the output word $p_g(x^{'}_t)$:

$$P_g(x^{'}_t) = softmax(V^{'}(V[s_t, c^t] + b) + b^{'}) \tag{3.8}$$

where $V^{'}$, V, b, and $b^{'}$ are learnable parameters. We finally use a copy mechanism as presented in the *Pointer Generator Model* (PGN) (See et al., 2017) to consider Out-Of-Vocabulary words. As detailed in (Bražinskas et al., 2019), it will also let the model access rare words from the group reviews and reduces out of context hallucinations. The new probability of generating the output word $x^{'}_t$ becomes:

$$p_{gen} = \sigma(W^\top_{Cgen} c^t + W_{Sgen} s_t + W_x x^{'}_{t-1} + b_{pgen}) \tag{3.9}$$

$$P(x^{'}_t) = p_{gen} \times P_g(x^{'}_t) + (1 - p_{gen}) \times \sum_{i \in V_{ext}} (a^t_{\_i}) \tag{3.10}$$

where W_{hgen}, W_{Sgen}, W_x, and b_{pgen} are learnable parameters, σ is the sigmoïd function, and V_{ext} is the extended vocabulary aggregating the vocabulary and the source document distributions. As can be seen, this approach allows direct access to the words of the reviews in the same group. The authors of *LSumm* explain that this technique avoids out of context hallucinations (Bražinskas et al., 2019) by sampling rare terms from group reviews. This still requires a certain homogeneity of the input texts so that the selected entities are coherent. To pursue and push forward this idea of controlled hallucinations, we let the model choose to also draw words directly from the distribution of topics $p(BoW_i^{'})$ defined in equation 3.3.1 by modifying the final probability of selecting a word:

$$P_{final}(x_t^{'}) = P(x_t^{'}) + p(BoW_i^{'}) \tag{3.11}$$

We notice empirically taht letting the model choose between the two probability helps the model to converge better when learning the topic distribution. The language model is trained with the cross-entropy function.

General Architecture

Our complete approach combines the topic and the language models in the objective is to maximize the following function:

$$\begin{aligned} \log \int \left[p_\theta(c) \prod_{i=1}^{M} \int p_\theta(R_i|z_i, R\_i, BoW_i, t_i) p_\theta(z_i|c) dz_i \right] dc \\ + \log \int \prod_{i=1}^{M} p_\theta(BoW_i|t_i, \beta) dt_i \end{aligned} \tag{3.12}$$

The right part of the function describes the topic model, and the left part depicts our language model conditioned by the topic content. This approach enables the system to learn relevant topics and use them to condition summary generation. More specifically, we can use the average of c, t, and the bag-of-word representation to generate a typical review and bias it toward the main topic of the group of documents.

3.3.2 Model distributions

In this section we describe in detail the set of assumptions about the prior and posterior distributions of the different latent variables c, z, and t and the methods used to sample them. As we use variational autoencoders for both the topic and the language models, we are going to approximate the posterior distributions with a different inference neural networks parametrized by Φ.

Group review latent variable: c

Once again, we observe the principles defined for Lsumm in (Bražinskas et al., 2019)by assuming a standard normal prior distribution $p(c) = \mathcal{N}(c;0,I)$. Following the reparameterization trick (Kingma and Welling, 2013) for Gaussian distribution, we use a linear projection on h_c specified in equation 3.3.1 to obtain the parameters of the inference posterior distribution:

$$\begin{aligned} \mu_\Phi(c) &= W_{meanc}H_c + b_{meanc} \\ \sigma_\Phi^2(c) &= exp(W_{varc}H_c + b_{varc}) \end{aligned} \tag{3.13}$$

where W_{meanc}, b_{meanc}, W_{varc}, b_{varc} are learnable parameters of the model, $\mu_\Phi(c)$ is the mean, and $\sigma_\Phi(c)$ is the variance of the distribution of the approximated inference network $q_\Phi(c|h_c) = \mathcal{N}(c;\mu_\Phi(h_c), I\sigma_\Phi(h_c))$.

Individual review latent variable: z

We also assume a normal Gaussian distribution for the prior of z. The major difference is that the latter is conditioned by the latent variable c to obtain $p_\theta(z|c) = N(z;\mu_\theta(c), I\sigma_\theta(c))$. Regarding the parameters of the inference posterior distribution, we use the same procedure by linearly projecting the concatenation $[R_i;c]$. Then, we estimate $\mu_\Phi(z)$ as the mean and $\sigma_\Phi^2(z)$ as the squared variance of $q_\Phi(z_i|R_i,c) = N(z_i;\mu_\Phi(R_i,c), I\sigma_\Phi(R_i,c))$.

Topic latent variable: t

For the latent topic variable we assume a Dirichlet prior distribution because it has been shown useful to obtain good and interpretable topics (Blei et al., 2003). The major issue is that the reparametrization trick is difficult to implement for this particular type of distribution. Therefore, we employ the methods introduced in *AVTIM* (Srivastava and Sutton, 2017) to approximate the distribution and make it tractable within the VAE framework. The authors resolve this issue by proposing a Laplace approximation with a softmax estimation. They demonstrate that this approximation to the Dirichlet prior $p_\theta(t|\alpha)$ is equivalent to consider t as a multivariate normal with mean $\mu(t)$ and covariance matrix $\sigma(t)$ that we can estimate with the defined α parameter vector of the Dirichlet distribution:

$$\mu_k(t) = \log(\alpha_k) - \frac{1}{K}\sum_i^K \alpha_i \tag{3.14}$$

$$\sigma_{kk}(t) = \frac{1}{\alpha}(1 - \frac{2}{K}) + \frac{1}{K^2}\sum_i^K \frac{1}{\alpha_k} \tag{3.15}$$

where K is the total number of topics and the dimension of the α vector defined as an hyperparameter for Dirichlet based models. This corresponds to consider the distribution of the topics over α as a logistic normal distribution with parameters μ and σ. Once we assume this distribution, we can compute the parameters of the posterior distribution from the inference network as a linear projection on h_i^{BoW} as described in equation 3.3.2 to obtain $q_\Phi(t_i|h_i^{BoW}) = N(t_i; \mu_\Phi(h_i^{BoW}), I\sigma_\Phi(h_i^{BoW}))$.

3.3.3 Model loss function

As we have seen so far, we thus have two models that are combined to fulfill a multi-task learning objective. For variational inference, such as Bayesian models or VAEs, the computation of the marginal likelihood requires exponential time as it needs to be evaluated over all configurations of latent variables. Therefore, we seek to maximize the Evidence Lower BOund (ELBO) regarding both the parameters θ and Φ. The following equations depict the language model noted $\mathscr{L}_{LM}$ and the topic model loss $\mathscr{L}_{TM}$.

$$\begin{aligned}\mathscr{L}_{LM}(\theta,\Phi) = \mathbb{E}_{q_\Phi(c|R)}\Bigg[\sum_{i=1}^{M}\mathbb{E}_{q_\Phi(z_i|R_i,c)}&\\ [\log p_\theta(R_i|z_i,t_i,BoW_i)] -&\\ \sum_{i}^{M}\mathbb{D}_{KL}[q_\Phi(z_i|R_i,c)||p_\theta(z_i|c)]\Bigg]&\\ -\mathbb{D}_{KL}[q_\Phi(c|R)||p_\theta(c)]&\end{aligned} \tag{3.16}$$

$$\begin{aligned}\mathscr{L}_{TM}(\theta,\Phi) = \sum_{i=1}^{M}\mathbb{E}_{q_\Phi(t_i|BoW_i)}[\log p_\theta(BoW_i|t_i,\beta)&\\ -\mathbb{D}_{KL}[q_\Phi(t_i|BoW_i)||p_\theta(t_i|\alpha)]]&\end{aligned} \tag{3.17}$$

For both losses, the left part of the expressions respectively ensure the text reconstruction of R_i or its bag of words representation BoW_i. The second term is the *Kullback-Leibler* divergence, which compels a learned posterior distribution to reach its respective prior distribution. For the Dirichlet and Gaussian distribution hypotheses, the KL divergence terms are computed thanks to the reparameterization trick described section 3.3.2 and detailed in the losses equation in (Bražinskas et al., 2019; Srivastava and Sutton, 2017). Finally, the total loss of our model is thus:

$$\mathscr{L}_{tot} = \mathscr{L}_{LM} + \mathscr{L}_{TM} \tag{3.18}$$

3.3.4 Summary Generation

Our objective being to condition the generation of summaries according to one or several topics, we must first set up a strategy to designate the $k = [1,...,K]$ main topic(s) to include. For generative topic models, an interesting and relevant topic deviates the most from its expected prior distribution (AlSumait et al., 2009). We sample the posterior distribution from the bag of words representation and the prior topic t_{prior} distribution from the prior mean $\mu(t)$ and variance $\sigma(t)$ learned during the training phase. To select the K main topics and ensure their diversity, we implemented a Maximum Margin Relevance approach (Carbonell and Goldstein, 1998). Therefore, we choose topics from the posterior distribution that maximize $cos(t_k^{prior}, t_k) - \lambda * cos(t_k, t_j)$, where t_j are the already picked topics and $\lambda = 0.5$.

For each topic t_k where $k = [1,...,K]$, to condition effectively the summary generation toward the topic, we bias the hidden representation h_c with the posterior topic-word distribution β. We select the vector β_k associated with the topic and we preserve $1/8$ of the most topically probable terms from the extended vocabulary. We tested multiple filtering factors ranging from $1/2$ to $1/32$. Our first observations let us think that if we keep too many words we do not impose enough diversity in the outputs, and if we remove too much then sentences become ungrammatical. We found that keeping $1/8$ words is a good compromise between the diversity and the coherency of the produced summaries. When creating h_c, instead of attending to all the group reviews' words, we attend only to X_{topics}, the set of topically relevant terms in the group reviews. The remaining words are masked, and equations 3.3.1 and 3.3.1 thus become:

$$\begin{gathered} a_{ij} = softmax(v^\top \tanh(W_c, m_{ij}, b_c)) \\ h_c = \sum_{i=1}^{M} \sum_{j=1}^{X_{topics}} a_{ij} \end{gathered} \tag{3.19}$$

Where $m_{ij} = [h_{ij}; E_{ij}]$ is the concatenation of the word $x_j \in X_{topics}$ embedding and GRU representations in the review R_i. Then we follow the recommendations prescribed in

LSumm (Bražinskas et al., 2019). We set c to $\mu_{\Phi}(c)$ constructed via the inference model through this biased h_c. Since the objective is to summarize a collection of reviews, we also start by setting z to its prior mean $z = \mu_{\theta}(c)$. However, we use the topic distribution per document t_k and to create $z^{topic} = \mu_{\theta}(c) * t_k$ a topically biased representation for each review depending on its topic correlation. We initialize the decoder to the concatenation $s_0 = [z; t_k]$. At this point, since we have defined these variables to their mean values, we end up with a one-dimensional vectors c and s_0 for the whole batch of reviews. We sample our summary by maximizing the probability expectation $P(x_t^{'})$ only. Contrarily to the training phase, we do not account for $p(BoW_i^{'})$ because we observe that directly biasing the probability with topical words extremely decrease the summary coherency. However, we have used $p(BoW_i^{'})$ in the beam search method to select among our K best hypothesis the one maximizes the sum of the two probabilities. Finally, we have also used $p(BoW_i^{'})$ in post-processing phase to replace the last unknown tokens generated by the model with the words that maximize it.

3.4 Experiments

3.4.1 Dataset

We trained our model on the Amazon Product dataset composed of reviews on 29 product categories (He and McAuley, 2016). We used similar pre-processing as in (Cho et al., 2014). We have considered products with a minimum of 15 reviews. We remove products with more than 350 reviews and after that, the ones above the 90^{th} percentile to reduce bias from specific terms associated with these items. We use the Spacy tokenizer [2] and we do not consider texts under 8 and above 200 tokens for identical reasons. For the evaluation, we employed the same 200 human-generated summaries as in (Bražinskas et al., 2019). The major difference is that, since we aim at improving the model ability to handle heterogeneous information, we sample our reviews from 19 categories which

[2] `https://spacy.io/api/tokenizer`

include the four used in the article (electronics, health care, home appliances, and clothes). Due to our hardware limitations, we decide not to utilize the full reviews remaining after filtering, and set our final training data to 17,497 reviews drawn from 303 products, and the validation to 3,105 reviews from 50 products.

3.4.2 Implementation

For our experiments, our model uses the GloVe 200 dimensional pre-trained word embeddings (version: glove.6B.200d) (Pennington et al., 2014). The text was lowercased and we utilized the Spacy tokenizer. Both the encoder and the decoder of the model are composed of a single bidirectional layer with a size of 512 hidden units. We set the dimensions of the latent variable z and c to 600 and the hidden layer or the c inference attention network at a dimension of 200. For the bag of words representation, we used the Spacy part-of-speech tagger and preserve only adverbs, adjectives and nouns. We set the number of topics, and thus the dimension to the latent variable t to 30. During training, we initialize the weight of the different layers via a Xavier uniform distribution (Glorot and Bengio, 2010), and we established the dropout of each layer at 0.2. We train the model for 250 epochs with the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-4} for the language model. We have used a weight decay of 8^{-3} and a gradient clipping of 10. We train the model with stochastic gradient descent using mini batches containing 8 reviews. Regarding the KL annealing issue, we have employed a cycling function with $r = 0.8$ (Fu et al., 2019) with a maximum value of 1 for z and 0.65 for c. We have used a linear scheduling function between epochs 0 to 40 with a max value set to 1 for t. Finally, to improve further the output results, we apply the beam search method with a beam size established to 5 and an n-gram blocking method (Paulus et al., 2017) set to avoid trigram repetitions. We implemented our model with the Pytorch library[3] version 1.8.1. The model was trained on a machine with an 8 Gb NVIDIA Tesla P4 graphics card and a 60Gb, 16-core processor. Both the dataset and our code are available on GitHub[4].

[3] `https://pytorch.org/`

[4] `https://github.com/fcarichon/TopicDiversifiedVAESumm`

3.4.3 Evaluation metrics

We use the evaluation dataset introduced in (Bražinskas et al., 2019) composed of 3 human-generated summaries for 60 products consisting of 8 reviews. We report the average and maximum ROUGE F1 scores of the different methods Lin (2004) for the different baselines on the evaluation dataset. We also provide the ROUGE scores with filtered stop words as well to emphasize better the presence of content words. Since our objective is also to highlight the benefice of our topically diversified summaries, we observe the BLEURT score (Sellam et al., 2020) between the product reviews and the generated summaries. BLEURT allows us to indicate to what extent the summary conveys the overall meaning of the input. Finally, following a similar purpose, we report additionally how well our model can capture the topics addressed in the reviews. To that extent, we train a LDA model with the Gensim library [5] on our training dataset. Then we generate a topic distribution for both our summaries and test reviews. We first analyze the cosine similarity between their respective topic distribution. But we also evaluate the overlap between the top 100 most probable words drawn from the main topic of both inputs. We ponder the overlap by their probability distribution in the topic to account more if two words have high a high probability to appear in their respective topics.

3.4.4 Baselines

We compare our model with 3 different baselines. Two extractive models that are known for their strong performance in general-purpose summarization task and our base abstractive model *Lsumm*.

BERT for Text Summarization (Miller, 2019): This model is considered as a state-of-the-art model for extractive summarization. The BERT model is a general language model based on the transformer architecture. It has been fine-tuned for summarization purpose. In this study we do not fine-tune the model on our dataset to come close to an unsupervised configuration.

[5] `https://radimrehurek.com/gensim/models/ldamodel.html`

TextRank (Mihalcea and Tarau, 2004): This model is a graph based oriented summarization method for unsupervised extractive summarization. The method demonstrated throughout the years it strong performances on multiple datasets and set-ups.

Lsumm (Bražinskas et al., 2019): The model is an unsupervised abstractive summarization model relying on a variational autoencoder. We trained and fine-tuned the model on our Amazon training dataset with the same parameters detailed in section 3.4.2. Since it is the basis of our approach, we consider this model as the vanilla version that ignores any topic information.

3.5 Results and experimental analysis

3.5.1 Model evaluation

We report the results from the baselines and from our approach with 2 different configurations. our first configuration *TopiCatSumm* matches the other methods by producing one summary for the batch of reviews. If we define N_{mean} as the average length of a reviews batch, then our summary must equal this length. Therefore we produce K topically conditioned sentences of length N_{mean}/K, and which we concatenate to create the final summary. Since the average length of a review is 58 words, we set $K = 3$ to produce enough diversity in the output while still generating long enough pieces of texts to be coherent. For the second configuration *TopicNSumm*, we have decided to match the evaluation dataset by also producing 3 topically different outputs of length N_{mean}.

Since we have 3 references per group of reviews, we report for all the metrics both the average score between the summary and all the references, and the maximum score with the best matching reference. In the case of our second configuration *TopicNSumm*, we use the fact of generating multiple summaries to optimize the ROUGE score. Therefore, we first pair each human reference with the generated output that maximizes its score and we then report the average and maximum for all associated metrics. The ROUGE results are presented in table 3.1.

Table 3.1: ROUGE scores on the Amazon dataset. R-1a, R-2a, R-La stand for averaged results between the summary and the 3 human references while R-1m, R-2m, R-Lm stand for the maximum results between the summary and the references.

Methods	R-1a	R-1m	R-2a	R-2m	R-La	R-Lm
BERT Summarizer	25.03	30.33	4.17	7.39	15.31	18.67
TextRank	29.42	34.87	5.1	8.36	16.82	20.17
LSumm	17.57	21.92	0.51	1.14	10.91	13.59
TopiCatSumm	16.91	20.32	0.34	8.83	9.75	11.79
TopicNSumm	19.64	23.24	0.78	1.9	11.58	13.9

Table 3.2: Topic content and coverage evaluation on Amazon dataset.

Methods	R-1 filt. average	R-1 filt. maximum	BLEURT	Word topic overlap	topic similarity
Human references	NA	NA	-0.464	1.10	0.488
BERT Summarizer	18.64	25.04	-0.774	0.469	0.336
TextRank	21.24	27.29	-0.673	0.521	0.338
LSumm	6.67	9.89	-0.889	0.201	0.2
TopiCatSumm	9.39	12.71	-0,579	0.494	0.383
TopicNSumm	11.58	15.53	-0,677	0.678	0.477

The analysis of the ROUGE scores reveals that the abstractive approaches perform significantly worse than the extractive ones. These results differ from those reported in (Bražinskas et al., 2019). We believe this represents the greater heterogeneity of the product categories used in the training of the two abstractive models, which does not impact the extractive methods. However, the outcomes also show that our model configuration *TopicNSumm* allows an efficient optimization for matching related summary with their reference. Our *TopiCatSumm* approach, concatenating 3 subtopics to form the summary, was the least effective in reproducing human text. We think this is partly due to the size constraint penalizing the production of coherent sequences but the topic diversity and content coverage is still more relevant. To confirm our point and offer a new angle of analysis, we present different results in the table 3.2. More specifically we report here the ROUGE scores with filtered stop words, the BLEURT scores and the topic metrics that we introduced section 3.4.3. We further exhibit the average score obtained by the 3 human references for these metrics when possible.

These different results reveal that both our approaches successfully cover more content from the original reviews and that information appears more related to the topic distribution than other systems. It also shows that our method improve our baseline abstractive model for capturing meaningful material used in the human references. The table 3.3 provides examples of generated texts by our model and the various baselines to give a view of these content differences.

However, all these observations show that we stay far from the extractive baselines on this point. Our main hypothesis is that the intrinsic homogeneity of a batch dealing with the same product enables these methods to remain relevant. Table 3.4 highlights compare the various models with a batch including 16 reviews from 2 different products. While we note a significant performance decrease for all baselines, especially when we look at content words, our approach seems to suffer less from increasing heterogeneity. For the abstractive model, we realize that it generates coherent sentences but that are dissociated from any products or aspects. Regarding, the extractive methods we can see a pure concatenation sequence from the two separated texts. With our model configuration *TopiCatSumm* we either detect the same concatenation of disconnected elements or the summary stresses only on one of the two product aspects. The second approach *TopicNSumm* become interesting at that point because the multiplication of outputs allows us to observe various contents focusing on common aspects, or having dedicated summaries for each product depending on the topic.

Once again, we further highlights these analyses by providing in table 3.5 one example of generated documents for 2 different products by our model and the various baselines.

With these different investigations, we conclude that our model achieve the goal of producing abstractive summaries that are topically oriented and diverse thus improving content relevancy and coverage for analyzing customer reviews. Since we have used *LSumm* as a foundation to study topic diversity in unsupervised summarization, the objective is not to resolve the core problems related to this baseline. We are aware of the coherency and capacity drawbacks implied by our architecture and we detail our choices in the limitation section of this article.

Table 3.3: Examples of generated texts - Part 1. Table with examples of generated texts. For each product we provide the text generated by our two configurations, the absractive model LSumm and the extractive model TextRank.

B0002U34HY (CHV1510 Vacuum filter)	Our model TopiCatSumm	Easy fix before expected not much monster filters but with regular use handles clean, seems sturdy. However this filter was difficult with product support, I read comparable CHV1510 on here as other. The dirty class hitting washable model construction of functionality CHV1510 ridiculous, quality functionality washable.
	Our model TopicNSumm summary 1	CHV1510 games was home from eating all i complained without such cool 3rd CHV1510 brand I and amount on them off position not one time with filter that are just guessing all color!
	Our model TopicNSumm summary 2	that said filter and cheaper on shipping as hair fast shipping here than what should is but for something changed after working. The filter holder showed that, what appears it properly had different place for filter like using generic brand at all!
	Our model TopicNSumm summary 3	For the fans mounted cold lights: positive copies filters the world has broken open when aid properly from CHV1510, so in some amounts source on wrench breaking during these are fantastic and I still recommend
	LSumm	it says harder. to install with filter as possible for filter! it takes some amounts. it seems too strong as opposed the original one of it and
	TextRank	This is the wrong filter if you are buying the CHV1510 Hand Vacuum. This item list listed with the vacuum – ’frequently bought together’ with the Black & Decker CHV9608 9.6 Volt Cyclonic-Action Cordless DustBuster BUT this filter does NOT fit!

Table 3.3: Examples of generated texts - Part 2. Table with examples of generated texts (Continued)

B0013EQ20Y (Frye Boots)	Our model TopiCatSumm	it wish my face soft hat, the boots it cozy lifts up nice. Comfy ugg Frye perfectly residue inside comfortable stretchy amounted just what the doctor ordered from boots all. I served comfy, boot though sticks right but quickly to safety snug evenly over, all socks together is
	Our model TopicNSumm summary 1	Indeed an excellent product and most excellent boots base and nice as wide in between all sizes up. It needs enough for all occasion beware of adjustments such all over cameras during!
	Our model TopicNSumm summary 2	Indeed comfy! Securely packaged, the it too and I am wearing! it makes great for heavy use thick rooms but tough construction and comfort, sound nicely tasted Frye but
	Our model TopicNSumm summary 3	comfy boots has already hanging down set I wish where had them on fire if there have many on bugs like paper itself while having. Overall this pair work well
	LSumm	it seems so sturdy enough like that is. it seems more sturdy than expected to get them again and was worth to try them! it seems more comfortable! it seems better with
	TextRank	they can be a beast to get on, like any boot fit to last; once on, they are incredibly comfortable. With a 20year break from not wearing Frye it was a pleasant surprise the quality has stood the test of time.

Table 3.4: Results on the Amazon dataset for 16 reviews. Evaluation of the various approaches summarizing batches of 16 reviews sampled from 2 different products categories of the Amazon dataset.

Methods	R-1 (avg)	R-1 (max)	R-1 filt. (avg)	R-1 filt. (max)
BERT Summarizer	18.63	25.04	13.35	21.96
TextRank	21.24	27.29	14.02	23.58
LSumm	18.39	25.58	4.13	6.17
TopiCatSumm	16.67	22.17	9.11	15.17
TopicNSumm	18.45	25.15	11.04	18.02

3.5.2 Model and configuration analysis

Since we introduce a modification of a preexisting model, we offer in this section a description and an analysis of our topic summarizer. During training, we have focused on understanding the integration of our topic model with our language and summarization model. First, the observation of the negative log likelihood (NLL) loss and KL divergence shows that, as indicated in (Xiao et al., 2018), it is essential to add the topic latent variable t for training the language model, and to combine it with the Bag of Words reconstruction function $matchalL_{TM}$. In our context of document summarization, there are two options for the Bag of Word representation. We can create a vector depicting each review individually, or a concatenation for the entire group. We can note from the study of NLL loss and the topic KL divergence that we require to keep an individual representation. If we use a representation for the whole group, then the model is unable to optimize both $\mathscr{L}_{TM}$ and $\mathscr{L}_{LM}$ at the same time. The need to capture individual and group information in the training either restrict too much or brings too much noise into the latent variable t, penalizing one loss or the other. Additionally, in our approach, we directly concatenate t with z to condition our decoder. However, since we employ the group representation c to generate our summary, it would be tempting to concatenate c and t to condition z and use only the latter in the decoder. We would then have the a priori distribution of z such that $p_\theta(z|[c,t]) = N(z;\mu_\theta([c,t]), I\sigma_\theta([c,t]))$, where $[c,t]$ is the concatenation of both variables. But once again we see a decrease in the NLL loss with this configuration, and

Table 3.5: Examples of generated texts for 16 reviews. Table with examples of generated text by the various models for batches of 16 reviews sampled from 2 different products of two different categories.

B0002U34HY vacuum filter & B00006IUVM kitchen steamer	Our model TopicNSumm summary 1	quality filters not do any reviews and picture looks as usual but for decades material seems fine but great purchase and deliver quality packaged! yeah and trust with
	Our model TopicNSumm summary 2	quality filter for many light steamer washable rice brand steamer, although is just easy enough without sending to play using without issues until much sized goes steamer easy too steam rice for each nut only goes straight smoothly
	Our model TopicNSumm summary 3	ladies! steam it has superior points of shelves from there : do something that? this steamer gives all aspects go, some kind opened without wearing them into this. So in some reviews from dragon appeared steam as directed, received mine ripped rice vegetables today
	LSumm	the filter is just what i needed. i have a lot of the filter and the filter. is not the same as the original filter.. is a great deal. is a great deal. is a great deal. is a very a very a very
	TextRank	This is the wrong filter if you are buying the CHV1510 Hand Vacuum. Sometimes I use the steamer for just one vegetable, or for rice, but it's really nice to have the separate basket.

we also note difficulties to train the topic KL divergence, resulting in low topic quality and inability to create diverse summaries. Our first hypothesis follows that of (Xiao et al., 2018), indicating that the application of the topic multimodal distribution facilitates the task of capturing information for z. Moreover, the use of the latent variable t in the model's early layers makes it possible to compensate learning and thus blur information in t with other hidden layers, such as the RNN or the attention in the decoder, not allowing stable training of the topic distribution.

During the summaries' construction, several options were possible and relevant to topically bias the text generation. We start by looking at the sampling of latent variable t representing the topic distribution for each document and the topic-by-word matrix generated through the β distribution. The tests carried out, and especially the study of the output texts, demonstrated the importance of generating these matrices from the posterior distribution. Indeed, the problem with using the prior distribution is that the topics are broad, having been learned from numerous heterogeneous sources. On the contrary, the posterior distribution produces summaries that are more relevant and include fewer hallucinations. The remaining options modify our current configuration described in section 3.3.4. Therefore, we iteratively customized several parameters and analyzed how they impacted the results:

- Configuration 0 (our current configuration): We use the matrix generated by the posterior β distribution to mask off-topic terms when creating the group representation h_c. We also bias the representation of z with the main topic distribution per document to obtain z^{topic}. Finally, we modify equation 3.3.1 for $P_{final}(x_t^{'}) = P(x_t^{'})$ not to account for the bag of words probability in text generation.

- Configuration 1: We sample t from its mean and variance, but as for c and t we could attempt to set it at its mean $mu(t)$ only.

- Configuration 2: We have set c to its posterior mean $mu_{\Phi}(c)$. However, we could also try to bias it with the main topic distribution by creating $c^{topic} = mu_{\Phi}^{topic}(c)$ as we do for z^{topic}.

- Configuration 3: Inversely, instead of employing the topic biased representation z^{topic}, we could set it to its mean $z = mu_{\theta}(z)$ as in *Lsumm*.

- Configuration 4: Rather than masking the attention related to the group review to generate h_c, we could weight the attention tensor with the word's topic probability.

- Configuration 5.a: Rather than masking attention h_c at the group level, we could mask attention used directly in the decoder. We would apply the same principle considered for equation 3.3.4 on equation 3.3.1.

Table 3.6: Configuration analyses. Table introducing the different results from various model configurations. We repeat the results of our main model in the first line for comparison.

TopicNSumm configurations	R-1 (avg)	R-1 (max)	BLEURT	Hidden diversity
Configuration 0	19.64	23.24	-0.677	0.578
Configuration 1	16.76	20.23	-0.656	0.513
Configuration 2	19.62	22.58	-0.69	0.534
Configuration 3	19.58	23.12	-0.65	0.328
Configuration 4	19.53	23.56	-0.72	0.469
Configuration 5.a	19.82	23.86	-0.63	0.557
Configuration 5.b	19.78	23.92	-0.61	0.558
Configuration 6	20.02	23.56	-0.66	0.57
Configuration 7	19.78	23.39	-0.677	0.562

- Configuration 5.b: As for configuration 4, we could also weight the decoder's attention tensor with the word's topic probability rather than masking it.
- Configuration 6: The copy mechanism described in equations 3.3.1 and 3.3.1 could be restricted to the identical mask, making only topically relevant terms accessible.
- Configuration 7: We could deploy the bag of words probability $p(BoW_i^{'})$ as in equation 3.3.1 for the summary generation.

For these analyses, we used our 3-summary strategy as it produced better outcomes. First, we report the ROUGE-1 and BLEURT results. We also add a diversity metric to emphasize issues met by some configurations. To that end, we re-use the trained encoder 3.3.1 to create a dense representation for each summary. We then measure the average cosine distance between these encodings. It thus highlights the diversity of information generated by our model. The table 3.6 displays the results obtained.

For configuration 1, we mainly observe a decrease in the ROUGE score. We believe this is because we have trained our language model to reconstruct reviews with a richer topic distribution, limiting our model's ability to generate suitable results with only the mean. The slight drop in diversity could also indicate a lesser ability to draw from the topic distribution effectively. If we now look at configuration 2, we see virtually no difference in performances from our current configuration. This nurtures our initial observations from the training loss regarding the lack of efficiency in biasing the group representation c, which must be compensated for by sampling z. The configuration 3 tends to emphasize this even more, since biasing z has a significant impact on the outputs' diversity. We also mention that this performance may further be attributed to the fact that the z bias lets us optimize the selection of results generated by the beam search because we can preserve the summary most closely related to the main topic. The configuration 4 also highlights the same phenomenon since we observe a decrease in both BLEURT and diversity when applying attention weighting instead of hard terms masking. We think that, in that stage of the model, implementing a soft bias on words is once again not enough to produce diverse outputs. Finally, regarding information filtering, we can note from configurations 5.a and 5.b that directly impacting the decoder is effective for generating relevant content, but the inspection of the summaries shows a huge loss in terms of coherence and readability. These results are corroborated by the analysis of configurations 6 and 7, where the same phenomenon is observed. This means that there is a need to find a trade-off between constraining the language model to increase the topically related and diverse information and its ability to coherent outputs.

Finally, an interesting configuration concerns topic selection. So far, we have picked the most important topics automatically, but one possibility is to let the user explore specific topics or aspects. In this context, the user defines a set of keywords $X^{user} = X_0^{user}, ..., X_U^{user}$ that they want to focus on when summarizing. We employ the topic-word matrix generated via posterior β distribution. We select the K main topics that maximize the probability $p(X^{user}|t_k)$ of generating the user input terms. We provide 3 examples in table 3.7 of summaries generated by inputting the word "price".

We can see with the products B0013EQ20Y and B00006IUVM that the model has indeed biased the summaries to include terms such as "expensive", "full cost", or even "budget" which connects to the price without necessarily directly referencing the latter. As we studied the original reviews, we notice that some of them contain price-related content. However, the B0002U34HY example, where no reviews mention information

Table 3.7: Examples of topically-biased generated texts. Examples of generated texts by our TopicNSumm where we input the word "price" to the model. The products presented here are the one used in the previous examples to allow comparison of output with this new bias.

Product	Generated summary
B0013EQ20Y (Frye Boots)	comfy noticeable! easy boots comfortable leather is inexpensive and wonderfully easy quality although is heavy as long to high although! instead i do wish that i have ordering it or worn on amazon.com since that it broke in two, only bought it 4 and times full cost
B00006IUVM (Kitchen Steamer)	updated hard 3 days! steam as use to force me rice is perfect with all customers at work budget is able with hesitant help at night supply store, too expensive than to sell items.
B0002U34HY (CHV1510 Vacuum filter)	CHV1510 filters is too and save dust the legs on top because occasionally leave volume under cycle i make sure look for washable filter or something. maybe it only keeps wet VF08

on the price, shows that the incentive might sometimes be limited, and that the model is unable to produce a text oriented toward that aspect. While this can be frustrating for the user, we think it is beneficial that the model does not hallucinate false material in this case.

3.5.3 Limitations and future research avenues

First, we have introduced many new hyperparameters and different configurations to obtain adequate results. In addition to those mentioned in the configuration analysis, we also had to make assumptions on the prior uniform Dirichlet distribution, on the Marginal Maximal Relevance (MMR) lambda parameter for topic choice, on the KL divergence parameter to limit gradient vanishing to learn topics, and on the window size to mask group information. But above all, the number of topics selected is a crucial hyperparameter that could have a huge influence on the quality of the topics extracted and will be dependent on the datasets studied.

The study of model configurations also highlighted another limitation of our approach. Indeed, we have shown that the balance between enforcing the inclusion of topic words and respecting a certain coherence in the texts produced is difficult to achieve. More generally, biasing language model can push it to predict terms that should have not been generated otherwise. Moreover, it can lead to an increase of word hallucinations, which are, of course, always interesting since they are linked to the topic, but which are not present in the original reviews.

Finally, we are aware of the limitations of our architecture based on single-layer RNNs, and that text coherency is far inferior to current models based on pretrained large language models (LLMs). However, we would first like to mention that we were confined by budget and access to sufficiently powerful machines to fine-tune them properly. In particular, the challenge of using these models would be to understand which layer should be biased with our method to obtain acceptable results, while ensuring that it is not absorbed by the capacity of these architectures. Studying simpler models then allows us to guarantee that our approach does indeed generate diversity in document summarization problems, and we leave the analysis of its application to LLMs for future work.

3.6 Conclusion to chapter 3

In this paper, we introduced a modification of an unsupervised topic or aspect-based method for multi-document summarization of product reviews. It relies on two variational autoencoders combined in a multitask learning objective. By explicitly accounting for topics when training a language model, our approach lets us produce diverse output for the same product input. Therefore, we can either optimize matching human-generated references, maximizing content coverage, or even focusing on specific aspects. As a result, the model improves the performances of its abstractive base structure for the Amazon dataset. We further showed the potential of topic identification to improve the management of heterogeneous data. Therefore, we plan to study in future work if it can help increase the capacity of unsupervised model for multi-document summarization. We also expect to assess if we could apply our method with more advanced architecture such as transformer models to strengthen its coherency. Meanwhile, we believe that in the framework of this study we have presented an interesting novel approach for biasing information in unsupervised models for summary diversification.

References

AlSumait, L., Barbará, D., Gentle, J., and Domeniconi, C. (2009). Topic significance ranking of lda generative models. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part I 20*, pages 67–82. Springer.

Amplayo, R. K., Angelidis, S., and Lapata, M. (2021). Aspect-controllable opinion summarization. *arXiv preprint arXiv:2109.03171*.

Angelidis, S., Amplayo, R. K., Suhara, Y., Wang, X., and Lapata, M. (2021). Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Angelidis, S. and Lapata, M. (2018). Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Arora, R. and Ravindran, B. (2008). Latent dirichlet allocation based multi-document summarization. In *Proceedings of the second workshop on Analytics for noisy unstructured text data*, pages 91–97.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. (2015). Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Bražinskas, A., Lapata, M., and Titov, I. (2019). Unsupervised opinion summarization as copycat-review generation. *arXiv preprint arXiv:1911.02247*.

Carbonell, J. and Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Chu, E. and Liu, P. (2019). Meansum: a neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*, pages 1223–1232. PMLR.

Coavoux, M., Elsahar, H., and Gallé, M. (2019). Unsupervised aspect-based multi-document abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 42–47.

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

Fang, H., Lu, W., Wu, F., Zhang, Y., Shang, X., Shao, J., and Zhuang, Y. (2015). Topic aspect-oriented summarization via group selection. *Neurocomputing*, 149:1613–1619.

Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., and Carin, L. (2019). Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*.

Ganesan, K., Zhai, C., and Han, J. (2010). Opinosis: A graph based approach to abstractive summarization of highly redundant opinions.

Gao, C. and Ren, J. (2019). A topic-driven language model for learning to generate diverse sentences. *Neurocomputing*, 333:374–380.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Gong, Y. and Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25.

He, R. and McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

Isonuma, M., Fujino, T., Mori, J., Matsuo, Y., and Sakata, I. (2017). Extractive summarization using multi-task learning with document classification. In *Proceedings of the 2017 Conference on empirical methods in natural language processing*, pages 2101–2110.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Li, X., Shen, Y.-D., Du, L., and Xiong, C.-Y. (2010). Exploiting novelty, coverage and balance for topic-focused multi-document summarization. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1765–1768.

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

Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Miller, D. (2019). Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Ozyurt, B. and Akcayol, M. A. (2021). A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: Ss-lda. *Expert Systems with Applications*, 168:114231.

Paulus, R., Xiong, C., and Socher, R. (2017). A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Pecar, S. (2018). Towards opinion summarization of customer reviews. In *Proceedings of ACL 2018, Student Research Workshop*, pages 1–8, Melbourne, Australia. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Ren, Z. and De Rijke, M. (2015). Summarizing contrastive themes via hierarchical non-parametric processes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 93–102.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Sellam, T., Das, D., and Parikh, A. P. (2020). Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Sohn, K., Lee, H., and Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.

Srivastava, A. and Sutton, C. (2017). Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Suhara, Y., Wang, X., Angelidis, S., and Tan, W.-C. (2020). Opiniondigest: A simple framework for opinion summarization. *arXiv preprint arXiv:2005.01901*.

Xiao, Y., Zhao, T., and Wang, W. Y. (2018). Dirichlet variational autoencoder for text modeling. *arXiv preprint arXiv:1811.00135*.

Yogatama, D., Liu, F., and Smith, N. A. (2015). Extractive summarization by maximizing semantic volume. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1961–1966.

Zhai, C., Cohen, W. W., and Lafferty, J. (2015). Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Acm sigir forum*, volume 49, pages 2–9. ACM New York, NY, USA.

General Conclusion and Discussion

The information relevance in a text mainly depends on the user it is aimed for. Indeed, the perception of the same content will be modified according to, for example, the user's prior knowledge, the set of documents they will have the opportunity to consult, and the depth of the information provided to them (Lavrenko, 2008). These influencing factors create a state of ambiguity regarding its need, which only a wide variety of proposals can effectively address (Clarke et al., 2008). A text corpus has various facets connected to it that can respond to user intention, such as temporality, semantics, emotions, message sender or receiver, or application domain influence. It is then possible to create biases in a model to address its various potential needs. This notion lies at the heart of all tasks that rely on understanding natural language, and automatic document summarization is no exception. Summarizing consists of establishing a short statement that abridges the information and reflects the gist of the discourse of single or multiple sources to produce a reduced version of the original material using a computer to complete a specific task for a precise audience (Mani, 2001). Each method to produce this text will then be composed of three main stages: representing or comprehending input information, evaluating or scoring the importance of information, and selecting the information to generate the output. Considering that the summary must address a specific task for a specific audience is tantamount to asserting that the perception of information relevance will depend on contextual factors. More explicitly, the notion of summary's usage and audience are the purpose factors that define the intention behind the model functioning (Hovy et al., 1999; Jones, 2007). This notion of relevance perception is fundamental for unsupervised document summarization since the design of the three main stages stems directly from the choice made by the algorithm's creator to meet these needs. However, the impact of this bias, mainly because of its multidimensional aspect, remains largely unexplored (Khosravani and Trabelsi, 2023). During this thesis, we have aimed to demonstrate the importance of these purpose factors and their link with model design to better respond to new tasks or users' needs. More specifically, we analyzed how information bias on multidimensional aspects such as temporality, sentiments, or semantics could respectively address the need for update, sentiment-oriented, or topic-diversified summarization. We carried out this study to understand and analyze different document summarization methods and show that datasets and performance evaluation metrics are vital in interpreting how these approaches work, as these contextual factors can bias them. Each paper in this thesis has pursued this objective:

- Our first paper introduced the update sentence compression task intended for news events with regular updates for people who would not have the opportunity to read news stories in their entirety and rely mainly on notifications. Our model is based on a representation and scoring function intimately linked to indicativity and news datasets. Moreover, the two first stages of the summarization process were modified to consider the requirements explicitly for novelty or consistency of information. We then suggested various metrics, such as the reuse of iterative terms, complemented by human analyses to confirm our model's ability to consider the specific issues related to update summarization.

- In our second paper, the objective was to extract useful information from customer reviews to improve the products and services of one company. While relevant for addressing frustrations, companies must also obtain constructive and coherent feedback from reviews to meet new demands. Therefore, we changed the point of view in the task of opinion summarization by regarding the companies as our target audience. The need for the consensual information of the main sentiment shifted into an objective, constructive summary. We have applied a method encapsulating central relevant information and modified the representation and scoring stages of the model to remove sentiment from the summary. To demonstrate the effectiveness of our approach, we once again proposed new metrics and an evaluation dataset to analyze the objectivity of a summary generated by such approaches.

- Finally, our 3rd study also explores customer reviews but considers the potential variety of topics and product aspects that can appear in this feedback. To enable users to explore data according to their needs, we designed a system that either provides a general view, maximizing topic coverage in a summary or generates a summary oriented to a specific aspect or keyword addressed in customer opinions. We modified the learning function of a generative model, enhancing information coverage so that it could focus on different topics. We then modified the summary to include as many aspects as possible or create distinct texts for a selected topic. Once again, we demonstrated the performance of our approach by completing the analysis with metrics related to the consistency and homogeneity of the topics' distribution covered in a summary.

With this work, we studied user tasks and needs related to different aspects of a text, such as novelty, cohesion, sentiment and diversity. Therefore, each model has contributed to answering our initial research question on the link between summarization goal factors and information bias in document summarization models. By playing on the multiplicity of audiences and uses, we emphasized the necessity of managing the degree of indicativity, informativeness, specificity, and genericity of information. Based on these observations, we have linked the description of the task and the audience to the definition of the information for influencing the three stages of summary generation to perform better than general models. To do this, we had to study how to use the notions of salience, representativeness, redundancy and relevance diversity to meet these tasks. Furthermore, each paper shows the importance of having dedicated datasets and metrics to observe these dimensions to evaluate these innovative designs properly. Finally, we have contributed to the literature and the natural language processing community by bringing a new angle of analysis that will continue improving the understanding of unsupervised document summarization.

Contributions

Several reviews are also interested in automatic document summarization models and study them from different angles, explicitly distinguishing these approaches by the methods employed (Gupta and Lehal, 2010; Lloret and Palomar, 2012; Ježek and Steinberger, 2008) or through the classical steps of representation, selection, and generating the summary (Allahyari et al., 2017; Ferreira et al., 2013). Apart from a very recent paper by Khosravani and Trabelsi (2023) that focuses on techniques too, and that confirms the growing interest in unsupervised approaches for document summarization, notably for their ability to be complemented by pre-trained large language models and to adapt quickly to various datasets, there is currently no systematic review dedicated to these methods, and that aims to improve understanding of their fundamental disparities. The first meaningful contribution of this thesis thus concerns the systematic analysis of unsupervised models for automatic text summarization, as well as the standard metrics and datasets from the very historical approaches to the very modern ones based on deep learning algorithms. More specifically, studies solely relying on algorithmic methods provide an interesting point of view to depict such systems. They often admit the limitations to explain core differences and to justify why some methods will perform better than others on the same dataset with standard metrics such as ROUGE (Lloret and Palomar, 2012; Khosravani and Trabelsi, 2023).

This work introduces a typology that takes its roots in the first analyses on the different dimensions characterizing a summary (Jones et al., 1999; Hovy et al., 1999). A system producing a general-purpose summary, even a very good one, cannot respond to all tasks and all user needs. There are structural factors linked to the task, the data, and the use of summaries which will influence the functioning of document summary systems and which can be grouped, once again, under three categories of factors: Input, Output, and Purpose factors. How we meet these specific factors will then impact the model's design. While input and output factors are visible and their influence on the models' behavior is easily observable, purpose factors are more challenging to consider, as they create an implicit link between input and summary. However, the choice of summary usage and target audience helps us understand why the same text segment may be perceived as relevant or not. This notion is rooted in the foundations of information theory and topical and novelty relevance definitions. It identifies the information to be presented and the contribution of knowledge it will make to the user (Lavrenko, 2008). By introducing this new typology, we offer a way to link unsupervised algorithms, datasets, and evaluation measures to the core definition of relevance and information theory. By highlighting this relationship with the stages of information representation, scoring, generation, and evaluation in unsupervised summarization approaches, we then provide a shift from methods to approaches, allowing us to determine and understand better some fundamental underlying concepts of information biases behind unsupervised text summarization (Peyrard, 2018; Jung et al., 2019).

Exploiting this original angle of analysis, we have been able to develop solutions better to address different contexts of applications related to the summarization task. In particular, we first addressed the specific task of update summarization. In this paper, we introduced a new model based on two competitive autoencoders that rely on a parameter to manage the novelty generated. We also proposed to measure novelty with a new human evaluation protocol and automated metrics based on word reuse. Our second paper focuses on an alternative audience, a company perspective, in opinion analysis. We have incorporated an adversarial approach to prevent a generalist model from capturing sentiment-related information, enabling us to obtain a more objective view and thus address our user's needs. In addition, we introduced a new dataset to evaluate this task and various metrics to measure objectivity, neutrality and subjectivity in our summaries. Finally, our last paper handled usage disparities for our summary by wielding generative topic models, allowing us to control the diversity of topics in user opinions. Once again, we proposed to analyze our data employing metrics based on topic distributions to gauge how information was embedded in summaries. Our second contribution is, therefore, to demonstrate concretely, through 3 examples, how to bias the information in the summary creation stages to respond to these factors. In addition, these examples have shown that if the analysis is limited to standard methods such as ROUGE, it is complicated to perceive the contribution of a model modification to answer a specific need. This work confirms the recommendations provided by Fabbri et al. (2021) on diversifying the tasks and methods for evaluating document summarization models to improve a deep understanding of their functioning.

These examples allow us to understand how the characterization of information has been modified during the different stages of summarization to meet these tasks. We arrive at our major contribution by combining these observations with the typology that our literature review has put forward. Indeed, this thesis confirms our initial intuitions about the link between certain aspects of document summarization and information theory. More specifically, the contextual purpose factors connect the information in the input document(s) to that included in the output (Jones et al., 1999). It is only natural to see these factors intimately tied to the very definition and perception of relevance (Peyrard, 2018). Two main characteristics materialize from contextual factors: the summary's usage and audience. When analyzing the relevance, two concepts emerge: topical and novelty relevance. Topical relevance marks the relevance related to the subject, linking the degree of thematic correspondence between the utilization need and the response received in a text. Novelty relevance identifies how the semantic content meets the user's need for information and complements their previous knowledge. A clear relationship can be seen between the summary's usage and topical relevance, where each definition is entwined in the context of application use for a produced output. The same obvious link can be made between novelty relevance and audience since both notions address the class of users targeted by the summary. Moreover, all these concepts can be concretely connected to the examination and discussion of automatic summarization approaches. Indeed, if we consider the summarization usage, we can distinguish two dimensions. Indicativity aims to promote content that enables understanding of a topic in detail, and informativeness seeks to describe what is being said and the overall content. When we analyze the methods, we observe a first category that expresses a document's specific information, spotlighting the input's characteristic elements that allow us to appreciate the topic and which is linked to this indicative dimension. The second category defines relevant information as central, letting us explore the maximum number of elements in the text, perfectly representing this notion of informativeness. Finally, we observe one interesting property of indicative and informative content in our classification, and that is linked to the work in (Mani, 2001): the fact that informative summaries can act as indicative ones, making the content of informative summaries a subset of the indicative ones. We can observe this in our approaches because characterizing representativeness still relies on statistical and topic properties, which are used in our topic selection category. This relation perfectly reproduces the subset relation between the purpose factors of a summary, and our classification of approaches makes this even more apparent. In the same way, the audience is described by two conceptions. Specificity seeks to bring the maximum amount of information on a subject to the user by filtering out useless information (Mani, 2001), and genericness tends to provide a complete and generalized view of the source material by covering as much of its information as possible (Mani, 2001). Once again, if we observe how methods encapsulate novel information, we observe the emergence of the first idea of non-redundancy, where we aim at maximizing the information gain of a specific theme. We see the notion of topic diversification, which seeks to maximize the coverage of the different topics addressed in a set of documents. These notions can subsequently be connected to specificity and genericness, respectively. Once again, we also note an interesting parallel between purpose factors and our distinction in the definition of novelty. Specific and generic summaries indeed cover complementary information. Both factors represent a spectrum where they are not incompatible. Systems can apply a combination of both redundancy and diversity to increase the covered content so long as the length constraints are respected (Huang et al., 2010; Chowdhury et al., 2021; Joshi et al., 2019). Our contribution to the literature becomes evident, as during this thesis, we provide further firm evidence for the link between purpose factors and automatic document summarization systems. Specifically, we have been able to demonstrate that the three stages of summarization all enable information to be represented, evaluated and produced in distinct formats to meet given tasks and users' needs, creating a concrete bridge to the very definition of the summarization task and the analysis and comprehension of computer-based methods that try to address this issue (Mani, 2001).

Discussion

How can we now interpret these results and contributions to analyze certain phenomena appearing in the current automatic document summarization literature? We have demonstrated the importance of considering the definition of usage, task, and users since they influence how each system will represent the information. Therefore, we can observe that not all systems are suited to all needs. Since the analysis of NLP models relies on comparison, choosing the right approaches for such a comparison is essential. Otherwise, by not considering this control factor, the analysis becomes easily open to criticism since it is impossible to say whether one model performs better than another and to understand the reasons for these differences in performance. We would also like to bring a new focus point specific to purpose and unsupervised methods. Naturally, humans produce better summaries after being trained to identify relevant source texts such textual features as topic sentences, keywords, and repeated ideas (Hidi and Anderson, 1986). Therefore, it is normal to notice the same phenomenon appearing in the summaries used by the automatic text summarization community. Specifically, several authors have observed that, when no instructions are specified, the human summaries, although different, are based on common properties such as employing term frequency (Nenkova and Vanderwende, 2005), including named entities, topic-specific terms (Delort and Alfonseca, 2012), and the noninclusion of reported facts and figures (Goldstein et al., 1999). Although many different guidance and tasks have been proposed at conferences such as DUC, NIST, or TAC, some have never defined these instructions, and others have been intentionally biased. In particular, the purpose, the intention, or the audience of the summary are never stipulated, which makes tasks always specific since the community was not satisfied with generic summaries given that they increased the variability of human experts' productions (Over et al., 2007). So, there is no reason to believe that, regarding these objectives, the human experts producing those outputs follow their natural tendencies, especially when we know that the most used data in the literature are news stories that tend to use events and named entities (Filatova and Hatzivassiloglou, 2004). Of course, one could argue that these issues are unimportant if good performance is achieved. However, other issues also arise in the context of performance measurement. We have assessed the current use of intrinsic evaluation techniques, especially the ROUGE evaluation method (Lin, 2004), which represents 70% of evaluation metrics used in the literature and is still used as the almost sole measure in recent works. The other main methods are the F1-score or the Pyramid method (Nenkova and Passonneau, 2004). These intrinsic metrics suffer from several flaws (Fabbri et al., 2021) such as some bias towards lexical similarity and do not account for fluency and readability (Scialom et al., 2019), or such as the fact that they are easy to fool and that one can obtain outstanding scores without producing a good summary because they rely highly on a frequency count where greedy methods can perform better than a consensus of human experts (Sjöbergh, 2007). The major issue related to the observations made in this thesis is that they all use reference summaries created by human experts, which will inevitably impact how they operate, with all the issues we have just raised concerning how these datasets are created. This phenomenon thus creates an implicit homogenization of the terms used to constitute the final document and, due to the nature of these elements, increases the possibility that our text is specific and indicative (Jones, 1972, 2007). The exclusive use of intrinsic performance metrics such as precision and recall or ROUGE, which is known to correlate very well with human production, thus favors the homogenization of the summary generated by automatic systems, as has already been observed in (Owczarzak and Dang, 2011). Given this new challenge, we can legitimately highlight the need to clarify these dimensions of purpose factors and propose performance measurement metrics independent of the dataset. If we stick to intrinsic measures, then it becomes vital to specify the conditions under which the datasets were created and to describe in detail how the relevant information was considered so that the appropriate methods and evaluations can be used. Moreover, this new practice will be useful for unsupervised methods and supervised approaches that rely even more heavily on data to function. On the other hand, because unsupervised approaches rely only on the implicit structure of the text and its underlying properties to identify the important elements to include in a summary, this brings them closer to the way humans summarize and to all issues coming from human summarization.

By spelling out all the conditions under which systems are created and evaluated, we then make them fit to be more suitable when there is no training or sparse data, domain, language, or field adaptations, or unknown conditions and external factors (Riedhammer et al., 2010). For all these situations, if human experts need to take time to digest all the information to create labels for each different situation, we are pulling in the opposite direction of the very first meaning of automatic text summarization. These conditions are often encountered in real-world applications and industries with much-specialized data and no gold standards available. Knowing the importance of summarizing documents, especially to help people better understand information, we hope this thesis will contribute to improving comprehension of unsupervised automatic systems and their functioning dimensions and, thus, bring them closer to positively impacting society.

Limitations

It is now essential to define the limits of our work and analysis, delineate the conceptual framework within which it is embedded, and outline the scope of our contributions. The main limitation stems from the fact that all our theory is drawn from observing the state of the art and the functioning of the models. We have set up the most exhaustive literature review possible on unsupervised methods. Through this review, we empirically examined existing papers and found this link between purpose factors and the behavior of different approaches. However, as we have already mentioned, our typology comes from a personal assertion and interpretation of the papers. Unfortunately, due to the lack of transparency on the purpose of the summarization, how the algorithm was built to meet that purpose, and above all the absence of specifications on the dataset and evaluation metrics, it was impossible to set up a coherent experimental protocol to validate our observations. Indeed, how can we implement metrics to evaluate the difference between centrality and selective salience without ourselves adding bias to these evaluations? Therefore, even if this thesis relays statements made by many other researchers, it is essential to clarify that our work contributes to the community by proposing a fresh point of view on document summarization systems but in no way constitutes a new theory.

The second limitation stems from our observation of our models' behavior. It directly impacts how purpose biases are considered in the information encoding design of specific systems. Indeed, our work aimed to demonstrate that it was possible to adapt generalist models to particular needs. We then omitted a whole range of other factors from the document summarization problem. As a result, we implemented models that forced information biases to be included in the generated texts and thus suffered from a loss of coherence and increased hallucinations. Moreover, by modifying existing algorithms, we often made them more unstable by adding numerous hyperparameters. Despite their usefulness in this thesis for demonstrating the importance of purpose factors in model design, it would be more appropriate to balance the set of all summary impact factors. To do this, we would need to develop systems that directly consider all these constraints or have methods that implicitly compensate for the factors left out. We are considering employing the new pre-trained LLMs to improve performances, especially textual coherence and fluency.

This brings us to the final limitation of our work concerning applying these new large language models. Our choice is explained by the complexity of these architectures and their training. Numerous studies have demonstrated that distinct layers encapsulate different types of information. How do you know where to implement an information bias to meet the specific needs of certain document summarization tasks? Furthermore, even more importantly, how to interpret the results, positive or not, obtained by these approaches. Especially when we know that many of these approaches have been pre-trained on datasets that will contain information biases that we cannot control and analyze. Therefore, we recognize that this considerably reduces the scope and applicability of our work. Further detailed investigations should be carried out to understand the possibility of biasing these models for different information factors. These experimental protocols must also consider the learning paradigm shift recently observed with new generalist models such as ChatGPT.

Future research

We propose to continue studying the applicability of information relevance characteristics within large language models. In future research, our approach would introduce an angle for the vaster goal of AI explainability. More specifically, in this framework, we could examine how to rework the absolute definition of salience for information capture (Bastings and Filippova, 2020), how information can modify prompt learning and thus the behavior of models (Ding and Koehn, 2021), and finally, how to complement methods for interpreting and analyzing results obtained in unique conditions (Wang et al., 2022; Jacovi et al., 2022). In the more specific context of automatic document summarization, understanding and exploring the differences in the behavior of pre-trained large language models would follow the recommendations made by Khosravani and Trabelsi (2023) to improve the potential use of these techniques. Once again, comprehending how to modify the information encoding to meet different needs is essential to making unsupervised approaches the most reliable for document summarization.

Of course, to be able to effectively analyze these notions and how they may be reflected in various models, we would further wish to continue our work by setting up a precise experimental protocol. Such a protocol would require explicitly defining the concepts of intention, purpose and type of text information. We therefore hope that future studies will first focus on developing datasets where reference labels are explicitly controlled. This also demands the creation of new performance metrics to differentiate between centrality and selectivity, while connecting them to human perception of indicativity and informativeness. This would allow us to complete our work to establish an accurate theory on the link between information and document summarization. It will further fulfill the needs put forward by Bhandari et al. (2020) on having alternative ways of understanding summarization approaches.

Finally, now that our comprehension of information lets us improve generalist methods for specific tasks, we hope to generalize our work to various application cases that have not yet been examined in our thesis. Knowing that there are distinct visions of information that we still need to address in this thesis, such as complexity, readability, inclusiveness, or partiality, we not only want to study how these metrics can be complementary to the notions of salience and novelty. We also envisage responding to various tasks such as surveys, e-mails, long stories, ultra-personalized summarization, or peculiar needs that have yet to be explored.

Conclusion

The emergence of the Internet has involved a large-scale digitization of classic communication networks, thereby creating a vast amount of available textual data. This quantity has become so substantial that it is now humanly impossible to handle and digest the existing information. The interest in automatic text summarization has become increasingly important in research and business communities. It has also established new opportunities and applications for the new data provided (email, scientific and medical papers, blogs, reviews), the recent possible tasks (update, sentiment-based, or personalized summarization), and the recent objectives they try to fulfill (answering questions, text overview, critics). Therefore, summarization systems are better understood and have seen significant improvements, mainly thanks to technological advances such as deep learning techniques, which have made such systems more than sentence-extraction systems. These improvements are noticeable for abstract summarization for both supervised and unsupervised approaches, which have become more consistent. With the ever-increasing digitalization of our communication media, it now seems essential that these new models address different needs to adapt to their users' growing needs. To this end, we considered it essential to return to the very foundations of summarization to understand how the different facets and aspects of information could influence the perceived relevance of the content produced. More concretely, this thesis then explored the design and use of unsupervised methods to better address the text summarization task. It highlights the relationship between information theory and users' needs according to characteristics beyond superficial textual features. It first provides a clear framework for understanding how information is selected in unsupervised models and how to build internal representations that allow it to complete their tasks. It also provides an original perspective for exploring external elements, such as evaluation metrics and connecting them to the human perception of information relevance. With the rise of the quality of textual production by large language models and general models such as ChatGPT, the potential for various applications become increasingly prominent, especially in text summarization for the industrial world. Therefore, it becomes even more relevant to understand how models capture relevance to propose systems that will answer these new needs.

Bibliography

Abdi, A., Shamsuddin, S. M., Hasan, S., and Piran, J. (2019a). Automatic sentiment-oriented summarization of multi-documents using soft computing. *Soft Computing*, 23(20):10551–10568.

Abdi, A., Shamsuddin, S. M., Hasan, S., and Piran, J. (2019b). Deep learning-based sentiment classification of evaluative text based on multi-feature fusion. *Information Processing & Management*, 56(4):1245–1259.

Abu-Jbara, A. and Radev, D. (2011). Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 500–509.

Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. (2009). Diversifying search results. In *Proceedings of the second ACM international conference on web search and data mining*, pages 5–14.

Ahmet, A. and Abdullah, T. (2020). Recent trends and advances in deep learning-based sentiment analysis. *Deep learning-based approaches for sentiment analysis*, pages 33–56.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Al-Radaideh, Q. A. and Bataineh, D. Q. (2018). A hybrid approach for arabic text summarization using domain knowledge and genetic algorithms. *Cognitive Computation*, 10(4):651–669.

Alami, N., En-nahnahi, N., Ouatik, S. A., and Meknassi, M. (2018). Using unsupervised deep learning for automatic summarization of arabic documents. *Arabian Journal for Science and Engineering*, 43(12):7803–7815.

Alguliev, R., Aliguliyev, R., et al. (2009). Evolutionary algorithm for extractive text summarization. *Intelligent Information Management*, 1(02):128.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268*.

Allan, J., Gupta, R., and Khandelwal, V. (2001). Temporal summaries of new topics. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 10–18.

Allan, J., Jin, H., Rajman, M., Wayne, C., Gildea, D., Lavrenko, V., Hoberman, R., and Caputo, D. (1999). Topic-based novelty detection: 1999 summer workshop at clsp, final report.

AlSumait, L., Barbará, D., Gentle, J., and Domeniconi, C. (2009). Topic significance ranking of lda generative models. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part I 20*, pages 67–82. Springer.

Amancio, D. R., Nunes, M. G., Oliveira Jr, O. N., and Costa, L. d. F. (2012). Extractive summarization using complex networks and syntactic dependency. *Physica A: Statistical Mechanics and its Applications*, 391(4):1855–1864.

Amini, M.-R. and Gallinari, P. (2001). Automatic text summarization using unsupervised and semi-supervised learning. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 16–28. Springer.

Amplayo, R. K., Angelidis, S., and Lapata, M. (2021a). Aspect-controllable opinion summarization. *arXiv preprint arXiv:2109.03171*.

Amplayo, R. K., Angelidis, S., and Lapata, M. (2021b). Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amplayo, R. K. and Lapata, M. (2020). Unsupervised opinion summarization with noising and denoising. *arXiv preprint arXiv:2004.10150*.

Angelidis, S., Amplayo, R. K., Suhara, Y., Wang, X., and Lapata, M. (2021). Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Angelidis, S. and Lapata, M. (2018a). Multiple instance learning networks for fine-grained sentiment analysis. *Transactions of the Association for Computational Linguistics*, 6:17–31.

Angelidis, S. and Lapata, M. (2018b). Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. *arXiv preprint arXiv:1808.08858*.

Angelidis, S. and Lapata, M. (2018c). Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Anuradha, G. and Varma, D. J. (2016). Fuzzy based summarization of product reviews for better analysis. *Indian Journal of Science and Technology*, 9(31):1–9.

Aone, C., Okurowski, M. E., and Gorlinsky, J. (1998). Trainable, scalable summarization using robust nlp and machine learning. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 62–66.

Arora, R. and Ravindran, B. (2008). Latent dirichlet allocation based multi-document summarization. In *Proceedings of the second workshop on Analytics for noisy unstructured text data*, pages 91–97.

Atkinson, J. and Munoz, R. (2013). Rhetorics-based multi-document summarization. *Expert Systems with Applications*, 40(11):4346–4352.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Banerjee, S., Mitra, P., and Sugiyama, K. (2015). Generating abstractive summaries from meeting transcripts. In *Proceedings of the 2015 ACM Symposium on Document Engineering*, pages 51–60.

Banerjee, S., Mitra, P., and Sugiyama, K. (2016). Multi-document abstractive summarization using ilp based multi-sentence compression. *arXiv preprint arXiv:1609.07034*.

Baron, R. A. (1993). Criticism (informal negative feedback) as a source of perceived unfairness in organizations: Effects, mechanisms, and countermeasures.

Barzilay, R. and Elhadad, M. (1999a). Using lexical chains for text summarization. *Advances in automatic text summarization*, pages 111–121.

Barzilay, R. and Elhadad, M. (1999b). Using lexical chains for text summarization. *Advances in automatic text summarization*, pages 111–121.

Barzilay, R. and Lee, L. (2004). Catching the drift: Probabilistic content models, with applications to generation and summarization. *arXiv preprint cs/0405039*.

Barzilay, R., McKeown, K., and Elhadad, M. (1999). Information fusion in the context of multi-document summarization. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 550–557.

Barzilay, R. and McKeown, K. R. (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

Bastings, J. and Filippova, K. (2020). The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? *arXiv preprint arXiv:2010.05607*.

Baziotis, C., Androutsopoulos, I., Konstas, I., and Potamianos, A. (2019). SEQˆ3: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression. In *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)*, pages 673–681, Minneapolis, Minnesota. Association for Computational Linguistics.

Belkin, N. J., Oddy, R. N., and Brooks, H. M. (1982). Ask for information retrieval: Part i. background and theory. *Journal of documentation*.

Bhandari, M., Gour, P., Ashfaq, A., Liu, P., and Neubig, G. (2020). Re-evaluating evaluation in text summarization. *arXiv preprint arXiv:2010.07100*.

Bing, L., Li, P., Liao, Y., Lam, W., Guo, W., and Passonneau, R. J. (2015). Abstractive multi-document summarization via phrase selection and merging. *arXiv preprint arXiv:1506.01597*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Blitzer, J., Dredze, M., and Pereira, F. (2007a). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

Blitzer, J., Dredze, M., and Pereira, F. (2007b). Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

Boudin, F., El-Bèze, M., and Torres-Moreno, J.-M. (2008). A scalable mmr approach to sentence scoring for multi-document update summarization. In *Coling 2008: Companion volume: Posters*, pages 23–26.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. (2015). Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Bražinskas, A., Lapata, M., and Titov, I. (2019). Unsupervised multi-document opinion summarization as copycat-review generation. *arXiv preprint arXiv:1911.02247*.

Bražinskas, A., Lapata, M., and Titov, I. (2020). Few-shot learning for opinion summarization. *arXiv preprint arXiv:2004.14884*.

Brunn, M., Chali, Y., and Pinchak, C. J. (2001). Text summarization using lexical chains. In *Proc. of Document Understanding Conference*. Citeseer.

Bysani, P. (2010). Detecting novelty in the context of progressive summarization. In *Proceedings of the NAACL HLT 2010 Student Research Workshop*, pages 13–18.

Cao, Z., Li, W., Li, S., and Wei, F. (2017). Improving multi-document summarization via text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Cao, Z., Wei, F., Dong, L., Li, S., and Zhou, M. (2015). Ranking with recursive neural networks and its application to multi-document summarization. In *AAAI*, pages 2153–2159. Citeseer.

Carbonell, J. and Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Cardenas, R., Galle, M., and Cohen, S. B. (2021). Unsupervised extractive summarization by human memory simulation. *arXiv preprint arXiv:2104.08392*.

Carichon, F., Fettu, F., and Caporossi, G. (2023). Unsupervised update summarization of news events. *Pattern Recognition*, 144:109839.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., et al. (2005). The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

Ceylan, H. and Mihalcea, R. (2009). The decomposition of human-written book summaries. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 582–593. Springer.

Chaturvedi, I., Cambria, E., Welsch, R. E., and Herrera, F. (2018a). Distinguishing between facts and opinions for sentiment analysis: Survey and challenges. *Information Fusion*, 44:65–77.

Chaturvedi, I., Ragusa, E., Gastaldo, P., Zunino, R., and Cambria, E. (2018b). Bayesian network based extreme learning machine for subjectivity detection. *Journal of The Franklin Institute*, 355(4):1780–1797.

Cheung, J. C. K. and Penn, G. (2014). Unsupervised sentence enhancement for automatic summarization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 775–786.

Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Chowdhury, R. R., Nayeem, M. T., Mim, T. T., Chowdhury, M. S. R., and Jannat, T. (2021). Unsupervised abstractive summarization of bengali text documents. *arXiv preprint arXiv:2102.04490*.

Chowdhury, S. B. R., Zhao, C., and Chaturvedi, S. (2022). Unsupervised extractive opinion summarization using sparse coding. *arXiv preprint arXiv:2203.07921*.

Christensen, J., Soderland, S., Etzioni, O., et al. (2013). Towards coherent multi-document summarization. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1173.

Christian, H., Agus, M. P., and Suhartono, D. (2016). Single document automatic text summarization using term frequency-inverse document frequency (tf-idf). *ComTech: Computer, Mathematics and Engineering Applications*, 7(4):285–294.

Chu, E. and Liu, P. (2019). Meansum: a neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*, pages 1223–1232. PMLR.

Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666.

Clarke, J. and Lapata, M. (2008). Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*, 31:399–429.

Coavoux, M., Elsahar, H., and Gallé, M. (2019a). Unsupervised aspect-based multi-document abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 42–47.

Coavoux, M., Elsahar, H., and Gallé, M. (2019b). Unsupervised aspect-based multi-document abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Colhon, M., Vlăduţescu, Ş., and Negrea, X. (2017). How objective a neutral word is? a neutrosophic approach for the objectivity degrees of neutral words. *Symmetry*, 9(11):280.

Conroy, J., Schlesinger, J. D., and O'leary, D. P. (2006). Topic-focused multi-document summarization using an approximate oracle score. In *Proceedings of the COLING/ACL 2006 main conference poster sessions*, pages 152–159.

Conroy, J. M. and O'leary, D. P. (2001). Text summarization via hidden markov models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 406–407.

Cristea, D., Postolache, O., and Pistol, I. (2005). Summarisation through discourse structure. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 632–644. Springer.

Dang, H. T. (2006). Duc 2005: Evaluation of question-focused summarization systems. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering*, pages 48–55.

Dang, H. T. and Owczarzak, K. (2008). Overview of the tac 2008 update summarization task. In *TAC*.

Delort, J.-Y. and Alfonseca, E. (2012). Dualsum: a topic-model based approach for update summarization. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 214–223.

Denil, M., Demiraj, A., and De Freitas, N. (2014). Extraction of salient sentences from labelled documents. *arXiv preprint arXiv:1412.6815*.

Dernoncourt, F., Ghassemi, M., and Chang, W. (2018). A repository of corpora for summarization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Deutsch, D. and Roth, D. (2020). Understanding the extent to which summarization evaluation metrics measure the information quality of summaries. *arXiv preprint arXiv:2010.12495*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dias, G. and Alves, E. (2005). Unsupervised topic segmentation based on word co-occurrence and multi-word units for text summarization. In *Proceedings of the ELECTRA Workshop associated to 28th ACM SIGIR Conference, Salvador, Brazil*, pages 41–48.

Ding, S. and Koehn, P. (2021). Evaluating saliency methods for neural language models. *arXiv preprint arXiv:2104.05824*.

Dohare, S., Karnick, H., and Gupta, V. (2017). Text summarization using abstract meaning representation. *arXiv preprint arXiv:1706.01678*.

Dong, Y., Mircea, A., and Cheung, J. C. (2020). Discourse-aware unsupervised summarization of long scientific documents. *arXiv preprint arXiv:2005.00513*.

Doran, W., Stokes, N., Carthy, J., and Dunnion, J. (2004). Assessing the impact of lexical chain scoring methods and sentence extraction schemes on summarization. In

International Conference on Intelligent Text Processing and Computational Linguistics, pages 627–635. Springer.

Dutta, M., Das, A. K., Mallick, C., Sarkar, A., and Das, A. K. (2019). A graph based approach on extractive summarization. In *Emerging Technologies in Data Mining and Information Security*, pages 179–187. Springer.

Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM (JACM)*, 16(2):264–285.

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.

Eyal, M., Baumel, T., and Elhadad, M. (2019). Question answering as an automatic evaluation metric for news article summarization. *arXiv preprint arXiv:1906.00318*.

Fabbri, A., Li, I., She, T., Li, S., and Radev, D. (2019a). Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Fabbri, A. R., Han, S., Li, H., Li, H., Ghazvininejad, M., Joty, S., Radev, D., and Mehdad, Y. (2020). Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. *arXiv preprint arXiv:2010.12836*.

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. (2021). Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

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

Fang, C., Mu, D., Deng, Z., and Wu, Z. (2017). Word-sentence co-ranking for automatic extractive text summarization. *Expert Systems with Applications*, 72:189–195.

Fang, H., Lu, W., Wu, F., Zhang, Y., Shang, X., Shao, J., and Zhuang, Y. (2015). Topic aspect-oriented summarization via group selection. *Neurocomputing*, 149:1613–1619.

Fang, X. and Zhan, J. (2015). Sentiment analysis using product review data. *Journal of Big Data*, 2(1):1–14.

Ferreira, R., de Souza Cabral, L., Freitas, F., Lins, R. D., de França Silva, G., Simske, S. J., and Favaro, L. (2014). A multi-document summarization system based on statistics and linguistic treatment. *Expert Systems with Applications*, 41(13):5780–5787.

Ferreira, R., de Souza Cabral, L., Lins, R. D., e Silva, G. P., Freitas, F., Cavalcanti, G. D., Lima, R., Simske, S. J., and Favaro, L. (2013). Assessing sentence scoring techniques for extractive text summarization. *Expert systems with applications*, 40(14):5755–5764.

Févry, T. and Phang, J. (2018). Unsupervised sentence compression using denoising auto-encoders. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 413–422, Brussels, Belgium. Association for Computational Linguistics.

Filatova, E. and Hatzivassiloglou, V. (2004). Event-based extractive summarization. *Proceedings of ACL Workshop on Summarization, volume 111.*

Filippova, K. (2010). Multi-sentence compression: Finding shortest paths in word graphs. In *Proceedings of the 23rd international conference on computational linguistics (Coling 2010)*, pages 322–330.

Frank, J. R., Kleiman-Weiner, M., Roberts, D. A., Niu, F., Zhang, C., Ré, C., and Soboroff, I. (2012). Building an entity-centric stream filtering test collection for trec 2012. Technical report, Massachusetts Inst of Tech Cambridge.

Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., and Carin, L. (2019). Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*.

Fu, X., Zhang, Y., Wang, T., Liu, X., Sun, C., and Yang, Z. (2021). Repsum: Unsupervised dialogue summarization based on replacement strategy. In *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)*, pages 6042–6051.

Ganesan, K., Zhai, C., and Han, J. (2010a). Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 340–348, Beijing, China. Coling 2010 Organizing Committee.

Ganesan, K., Zhai, C., and Han, J. (2010b). Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. *Proceedings of the 23rd International Conference on Computational Linguistics, pages 340–348. Association for Computational Linguistics.*

Ganesan, K., Zhai, C., and Viegas, E. (2012). Micropinion generation: an unsupervised approach to generating ultra-concise summaries of opinions. In *Proceedings of the 21st international conference on World Wide Web*, pages 869–878.

Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Gao, C. and Ren, J. (2019). A topic-driven language model for learning to generate diverse sentences. *Neurocomputing*, 333:374–380.

Gao, Y., Zhao, W., and Eger, S. (2020). SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 1347–1354, Online. Association for Computational Linguistics.

García-Hernández, R. A., Montiel, R., Ledeneva, Y., Rendón, E., Gelbukh, A., and Cruz, R. (2008). Text summarization by sentence extraction using unsupervised learning. In *Mexican International Conference on Artificial Intelligence*, pages 133–143. Springer.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–574.

Ghadimi, A. and Beigy, H. (2020). Deep submodular network: An application to multi-document summarization. *Expert Systems with Applications*, 152:113392.

Ghadimi, A. and Beigy, H. (2022). Hybrid multi-document summarization using pre-trained language models. *Expert Systems with Applications*, 192:116292.

Gillick, D. and Favre, B. (2009). A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pages 10–18.

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

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Goldstein, J., Kantrowitz, M., Mittal, V., and Carbonell, J. (1999). Summarizing text documents: Sentence selection and evaluation metrics. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 121–128.

Goldstein, J., Mittal, V. O., Carbonell, J. G., and Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. In *NAACL-ANLP 2000 Workshop: Automatic Summarization*.

Gong, Y. and Liu, X. (2001a). Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25.

Gong, Y. and Liu, X. (2001b). Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25.

Grusky, M., Naaman, M., and Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

Gui, L., Jia, L., Zhou, J., Xu, R., and He, Y. (2020). Multi-task learning with mutual learning for joint sentiment classification and topic detection. *IEEE Transactions on Knowledge and Data Engineering*.

Gupta, P., Pendluri, V. S., and Vats, I. (2011). Summarizing text by ranking text units according to shallow linguistic features. In *13th International Conference on Advanced Communication Technology (ICACT2011)*, pages 1620–1625. IEEE.

Gupta, V. and Lehal, G. S. (2010). A survey of text summarization extractive techniques. *Journal of emerging technologies in web intelligence*, 2(3):258–268.

Harabagiu, S., Hickl, A., and Lacatusu, F. (2007). Satisfying information needs with multi-document summaries. *Information Processing & Management*, 43(6):1619–1642.

Harabagiu, S. and Lacatusu, F. (2005). Topic themes for multi-document summarization. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 202–209.

Hatzivassiloglou, V., Klavans, J. L., Holcombe, M. L., Barzilay, R., Kan, M.-Y., and McKeown, K. (2001). Simfinder: A flexible clustering tool for summarization. *Proceedings of the Workshop on Summarization in NAACL-01. 2001.*

Havaei, M., Mao, X., Wang, Y., and Lao, Q. (2021). Conditional generation of medical images via disentangled adversarial inference. *Medical Image Analysis*, 72:102106.

He, R. and McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

Hendrickx, I., Daelemans, W., Marsi, E., and Krahmer, E. (2009). Reducing redundancy in multi-document summarization using lexical semantic similarity. In *Proceedings of the 2009 Workshop on Language Generation and Summarisation (UCNLG+ Sum 2009)*, pages 63–66.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Heu, J.-U., Qasim, I., and Lee, D.-H. (2015). Fodosu: multi-document summarization exploiting semantic analysis based on social folksonomy. *Information processing & management*, 51(1):212–225.

Hidi, S. and Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of educational research*, 56(4):473–493.

Hill, M. (1991). Writing summaries promotes thinking and learning across the curriculum: But why are they so difficult to write? *Journal of reading*, 34(7):536–539.

Hou, S. and Lu, R. (2020). Knowledge-guided unsupervised rhetorical parsing for text summarization. *Information Systems*, 94:101615.

Hovy, E., Lin, C.-Y., et al. (1999). Automated text summarization in summarist. *Advances in automatic text summarization*, 14:81–94.

Hovy, E. H., Lin, C.-Y., Zhou, L., and Fukumoto, J. (2006). Automated summarization evaluation with basic elements. In *LREC*, volume 6, pages 899–902. Citeseer.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Huang, L., He, Y., Wei, F., and Li, W. (2010). Modeling document summarization as multi-objective optimization. In *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, pages 382–386. IEEE.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Isonuma, M., Fujino, T., Mori, J., Matsuo, Y., and Sakata, I. (2017). Extractive summarization using multi-task learning with document classification. In *Proceedings of the 2017 Conference on empirical methods in natural language processing*, pages 2101–2110.

Jacovi, A., Bastings, J., Gehrmann, S., Goldberg, Y., and Filippova, K. (2022). Diagnosing ai explanation methods with folk concepts of behavior. *arXiv preprint arXiv:2201.11239*.

Jang, E., Gu, S., and Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., et al. (2003). The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Ježek, K. and Steinberger, J. (2008). Automatic text summarization (the state of the art 2007 and new challenges). In *Proceedings of Znalosti*, pages 1–12. Citeseer.

Jiang, Y., Finegan-Dollak, C., Kummerfeld, J. K., and Lasecki, W. (2018). Effective crowdsourcing for a new type of summarization task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 628–633.

Jin, F., Huang, M., and Zhu, X. (2010). A comparative study on ranking and selection strategies for multi-document summarization. In *Coling 2010: Posters*, pages 525–533.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Jones, K. S. (2007). Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449–1481.

Jones, K. S. et al. (1999). Automatic summarizing: factors and directions. *Advances in automatic text summarization*, pages 1–12.

Joshi, A., Fidalgo, E., Alegre, E., and Fernández-Robles, L. (2019). Summcoder: An unsupervised framework for extractive text summarization based on deep auto-encoders. *Expert Systems with Applications*, 129:200–215.

Jung, T., Kang, D., Mentch, L., and Hovy, E. (2019). Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. *arXiv preprint arXiv:1908.11723*.

Kågebäck, M., Mogren, O., Tahmasebi, N., and Dubhashi, D. (2014). Extractive summarization using continuous vector space models. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 31–39.

Kedzie, C., McKeown, K., and Diaz, F. (2015). Predicting salient updates for disaster summarization. In *Proceedings of the 53rd Annual Meeting of the Association for Com-*

putational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1608–1617.

Khoo, C. S. and Johnkhan, S. B. (2018). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4):491–511.

Khosravani, M. and Trabelsi, A. (2023). Recent trends in unsupervised summarization. *arXiv preprint arXiv:2305.11231*.

Kim, G. and Ko, Y. (2021). Effective fake news detection using graph and summarization techniques. *Pattern Recognition Letters*, 151:135–139.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kipfelsberger, P., Herhausen, D., and Bruch, H. (2016). How and when customer feedback influences organizational health. *Journal of Managerial Psychology*, 31(2):624–640.

Kobayashi, H., Noguchi, M., and Yatsuka, T. (2015). Summarization based on embedding distributions. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1984–1989.

Krishnan, J., Purohit, H., and Rangwala, H. (2020). Unsupervised and interpretable domain adaptation to rapidly filter tweets for emergency services. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 409–416. IEEE.

Kryściński, W., Keskar, N. S., McCann, B., Xiong, C., and Socher, R. (2019). Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

Kupiec, J., Pedersen, J., and Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 68–73.

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Laban, P., Hsi, A., Canny, J., and Hearst, M. A. (2021). The summary loop: Learning to write abstractive summaries without examples. *arXiv preprint arXiv:2105.05361*.

Lamsiyah, S., El Mahdaouy, A., El Alaoui, S. O., and Espinasse, B. (2021a). Unsupervised query-focused multi-document summarization based on transfer learning from sentence embedding models, bm25 model, and maximal marginal relevance criterion. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–18.

Lamsiyah, S., El Mahdaouy, A., Espinasse, B., and Ouatik, S. E. A. (2021b). An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings. *Expert Systems with Applications*, 167:114152.

Lavrenko, V. (2008). *A generative theory of relevance*, volume 26. Springer Science & Business Media.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.

Ledeneva, Y., Gelbukh, A., and García-Hernández, R. A. (2008). Terms derived from frequent sequences for extractive text summarization. In *International conference on intelligent text processing and computational linguistics*, pages 593–604. Springer.

Lee, J.-H., Park, S., Ahn, C.-M., and Kim, D. (2009). Automatic generic document summarization based on non-negative matrix factorization. *Information Processing & Management*, 45(1):20–34.

Leite, D. S., Rino, L. H., Pardo, T. A., and Nunes, M. d. G. V. (2007). Extractive automatic summarization: Does more linguistic knowledge make a difference? In *Proceedings of the Second Workshop on TextGraphs: Graph-based Algorithms for Natural Language Processing*, pages 17–24.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (2016). A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Li, P., Ye, Q., Zhang, L., Yuan, L., Xu, X., and Shao, L. (2021). Exploring global diverse attention via pairwise temporal relation for video summarization. *Pattern Recognition*, 111:107677.

Li, X., Shen, Y.-D., Du, L., and Xiong, C.-Y. (2010). Exploiting novelty, coverage and balance for topic-focused multi-document summarization. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1765–1768.

Liang, X., Wu, S., Li, M., and Li, Z. (2021). Improving unsupervised extractive summarization with facet-aware modeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1685–1697.

Lin, C., Ouyang, Z., Wang, X., Li, H., and Huang, Z. (2021). Preserve integrity in realtime event summarization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(3):1–29.

Lin, C.-Y. (2001). See-summary evaluation environment. *WWW site, URL: http://www. isi. edu/cyl/SEE*.

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

Lin, C.-Y. and Hovy, E. (2000). The automated acquisition of topic signatures for text summarization. In *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*.

Lin, H. and Bilmes, J. (2010). Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 912–920.

Lin, H. and Bilmes, J. (2011). A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 510–520.

Lin, H., Bilmes, J., and Xie, S. (2009). Graph-based submodular selection for extractive summarization. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 381–386. IEEE.

Liu, D., Wang, Y., Liu, C., and Wang, Z. (2006). Multiple documents summarization based on genetic algorithm. In *International Conference on Fuzzy Systems and Knowledge Discovery*, pages 355–364. Springer.

Liu, H., Yu, H., and Deng, Z.-H. (2015). Multi-document summarization based on two-level sparse representation model. In *Twenty-ninth AAAI conference on artificial intelligence*.

Liu, J., Cao, Y., Lin, C.-Y., Huang, Y., and Zhou, M. (2007). Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 Joint Conference on Em-*

pirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 334–342, Prague, Czech Republic. Association for Computational Linguistics.

Liu, P., Huang, C., and Mou, L. (2022). Learning non-autoregressive models from search for unsupervised sentence summarization. *arXiv preprint arXiv:2205.14521*.

Liu, Y., Zhong, S.-h., and Li, W. (2012). Query-oriented multi-document summarization via unsupervised deep learning. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Lloret, E., Ferrández, O., Munoz, R., and Palomar, M. (2008). A text summarization approach under the influence of textual entailment. In *NLPCS*, pages 22–31.

Lloret, E. and Palomar, M. (2012). Text summarisation in progress: a literature review. *Artificial Intelligence Review*, 37(1):1–41.

Lloret, E. and Palomar, M. (2013). Tackling redundancy in text summarization through different levels of language analysis. *Computer Standards & Interfaces*, 35(5):507–518.

Louis, A. and Nenkova, A. (2013). Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

Lovinger, J., Valova, I., and Clough, C. (2019). Gist: General integrated summarization of text and reviews. *Soft Computing*, 23:1589–1601.

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

Lynn, H. M., Choi, C., and Kim, P. (2018). An improved method of automatic text summarization for web contents using lexical chain with semantic-related terms. *Soft Computing*, 22(12):4013–4023.

Ma, S., Deng, Z.-H., and Yang, Y. (2016). An unsupervised multi-document summarization framework based on neural document model. In *Proceedings of COLING 2016, the*

26th International Conference on Computational Linguistics: Technical Papers, pages 1514–1523.

Malireddy, C., Maniar, T., and Shrivastava, M. (2020). Scar: sentence compression using autoencoders for reconstruction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 88–94.

Mani, I. (2001a). *Automatic summarization*, volume 3. John Benjamins Publishing.

Mani, I. (2001b). Summarization evaluation: An overview. *Proceedings of the NAACL 2001 Workshop on Automatic Summarization*.

Mani, I., House, D., Klein, G., Hirschman, L., Firmin, T., and Sundheim, B. M. (1999). The tipster summac text summarization evaluation. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*.

Marcu, D. (1997). From discourse structures to text summaries. In *Intelligent Scalable Text Summarization*.

Maybury, M. T. (1995). Generating summaries from event data. *Information Processing & Management*, 31(5):735–751.

McAuley, J., Targett, C., Shi, Q., and Van Den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.

McCreadie, R., Macdonald, C., and Ounis, I. (2014). Incremental update summarization: Adaptive sentence selection based on prevalence and novelty. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 301–310.

McCreadie, R., Santos, R. L., Macdonald, C., and Ounis, I. (2018). Explicit diversification of event aspects for temporal summarization. *ACM Transactions on Information Systems (TOIS)*, 36(3):1–31.

McDonald, R. (2007). A study of global inference algorithms in multi-document summarization. In *European Conference on Information Retrieval*, pages 557–564. Springer.

Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Mei, Q., Guo, J., and Radev, D. (2010). Divrank: the interplay of prestige and diversity in information networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1009–1018.

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

Mihalcea, R. and Tarau, P. (2004b). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Miller, D. (2019). Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Mnasri, M., de Chalendar, G., and Ferret, O. (2017). Taking into account inter-sentence similarity for update summarization. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 204–209.

Mohammed, M. B. and Al-Hameed, W. (2021). Cohesive summary extraction from multi-document based on artificial neural network. In *2021 7th International Conference on Contemporary Information Technology and Mathematics (ICCITM)*, pages 81–87. IEEE.

Mori, T., Nozawa, M., and Asada, Y. (2005). Multi-answer-focused multi-document summarization using a question-answering engine. *ACM Transactions on Asian Language Information Processing (TALIP)*, 4(3):305–320.

Nallapati, R., Zhai, F., and Zhou, B. (2017). Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Narayan, S., Vlachos, A., et al. (2019). Highres: Highlight-based reference-less evaluation of summarization. *arXiv preprint arXiv:1906.01361*.

Nenkova, A. and McKeown, K. (2012). A survey of text summarization techniques. In *Mining text data*, pages 43–76. Springer.

Nenkova, A. and Passonneau, R. J. (2004). Evaluating content selection in summarization: The pyramid method. In *Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl 2004*, pages 145–152.

Nenkova, A. and Vanderwende, L. (2005). The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005*, 101.

Nomoto, T. and Matsumoto, Y. (2001a). An experimental comparison of supervised and unsupervised approaches to text summarization. In *Proceedings 2001 IEEE International Conference on Data Mining*, pages 630–632. IEEE.

Nomoto, T. and Matsumoto, Y. (2001b). A new approach to unsupervised text summarization. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 26–34.

Nomoto, T. and Matsumoto, Y. (2003). The diversity-based approach to open-domain text summarization. *Information processing & management*, 39(3):363–389.

Oliveira, H., Ferreira, R., Lima, R., Lins, R. D., Freitas, F., Riss, M., and Simske, S. J. (2016). Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization. *Expert Systems with Applications*, 65:68–86.

Ono, K., Sumita, K., Research, S. M., Center, D., Komukai-Toshiba-cho, T. C., et al. (1994). Abstract generation based on rhetorical structure extraction. *arXiv preprint cmp-lg/9411023*.

Oved, N. and Levy, R. (2021). Pass: Perturb-and-select summarizer for product reviews. In *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)*, pages 351–365.

Over, P., Dang, H., and Harman, D. (2007). Duc in context. *Information Processing & Management*, 43(6):1506–1520.

Owczarzak, K. and Dang, H. T. (2011). Who wrote what where: Analyzing the content of human and automatic summaries. In *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, pages 25–32.

Ozyurt, B. and Akcayol, M. A. (2021). A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: Ss-lda. *Expert Systems with Applications*, 168:114231.

O'Donnell, M. (1997). Variable-length on-line document generation. In *the Proceedings of the 6th European Workshop on Natural Language Generation, Gerhard-Mercator University, Duisburg, Germany*.

Padmakumar, A. and Saran, A. (2016). Unsupervised text summarization using sentence embeddings. *Technical Report, University of Texas at Austin*, pages 1–9.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Pardo, T. A. S. and Rino, L. H. M. (2003). Temario: a corpus for automatic text summarization. Technical report, NILC Tech. Report NILC-TR-03-09.

Parveen, D., Ramsl, H.-M., and Strube, M. (2015). Topical coherence for graph-based extractive summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1949–1954.

Paulus, R., Xiong, C., and Socher, R. (2017). A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Pecar, S. (2018). Towards opinion summarization of customer reviews. In *Proceedings of ACL 2018, Student Research Workshop*, pages 1–8, Melbourne, Australia. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Peyrard, M. (2018). A simple theoretical model of importance for summarization. *arXiv preprint arXiv:1801.08991*.

Peyrard, M. and Eckle-Kohler, J. (2016). A general optimization framework for multi-document summarization using genetic algorithms and swarm intelligence. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 247–257.

Prabhumoye, S., Quirk, C., and Galley, M. (2019). Towards content transfer through grounded text generation. In *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)*, pages 2622–2632, Minneapolis, Minnesota. Association for Computational Linguistics.

Qazvinian, V., Radev, D. R., Mohammad, S. M., Dorr, B., Zajic, D., Whidby, M., and Moon, T. (2013). Generating extractive summaries of scientific paradigms. *Journal of Artificial Intelligence Research*, 46:165–201.

Radev, D. (2000). A common theory of information fusion from multiple text sources step one: cross-document structure. In *1st SIGdial workshop on Discourse and dialogue*, pages 74–83.

Radev, D. R., Allison, T., Blair-Goldensohn, S., Blitzer, J., Celebi, A., Dimitrov, S., Drabek, E., Hakim, A., Lam, W., Liu, D., et al. (2004a). Mead-a platform for multidocument multilingual text summarization.

Radev, D. R., Blair-Goldensohn, S., and Zhang, Z. (2001). Experiments in single and multidocument summarization using mead. In *First document understanding conference*, page 1À8. Citeseer.

Radev, D. R., Jing, H., Styś, M., and Tam, D. (2004b). Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):919–938.

Radev, D. R. and Tam, D. (2003). Summarization evaluation using relative utility. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 508–511.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Rankel, P. A., Conroy, J., Dang, H. T., and Nenkova, A. (2013). A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 131–136.

Reichheld, F. (2006). The ultimate question: Driving good profits and true growth. *Boston, MA*.

Ren, Z. and De Rijke, M. (2015). Summarizing contrastive themes via hierarchical non-parametric processes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 93–102.

Ribeiro, R. and de Matos, D. M. (2007). Extractive summarization of broadcast news: Comparing strategies for european portuguese. In *International Conference on Text, Speech and Dialogue*, pages 115–122. Springer.

Ribeiro, R. and de Matos, D. M. (2011). Centrality-as-relevance: support sets and similarity as geometric proximity. *Journal of Artificial Intelligence Research*, 42:275–308.

Riedhammer, K., Favre, B., and Hakkani-Tür, D. (2010). Long story short–global unsupervised models for keyphrase based meeting summarization. *Speech Communication*, 52(10):801–815.

Riedhammer, K., Gillick, D., Favre, B., and Hakkani-Tür, D. (2008). Packing the meeting summarization knapsack. In *Ninth Annual Conference of the International Speech Communication Association*.

Rossiello, G., Basile, P., and Semeraro, G. (2017). Centroid-based text summarization through compositionality of word embeddings. In *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*, pages 12–21.

Rothe, S., Maynez, J., and Narayan, S. (2021). A thorough evaluation of task-specific pretraining for summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 140–145.

Rudra, K., Ganguly, N., Goyal, P., and Ghosh, S. (2018). Extracting and summarizing situational information from the twitter social media during disasters. *ACM Transactions on the Web (TWEB)*, 12(3):1–35.

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Saggion, H. and Gaizauskas, R. (2004). Multi-document summarization by cluster/profile relevance and redundancy removal. In *Proceedings of the Document Understanding Conference*, pages 6–7.

Sankarasubramaniam, Y., Ramanathan, K., and Ghosh, S. (2014). Text summarization using wikipedia. *Information Processing & Management*, 50(3):443–461.

Schiffman, B., Nenkova, A., and McKeown, K. (2002). Experiments in multidocument summarization. *Proceedings of HLT*, pages 52——58.

Schumann, R. (2018). Unsupervised abstractive sentence summarization using length controlled variational autoencoder. *arXiv preprint arXiv:1809.05233*.

Schumann, R., Mou, L., Lu, Y., Vechtomova, O., and Markert, K. (2020). Discrete optimization for unsupervised sentence summarization with word-level extraction. *arXiv preprint arXiv:2005.01791*.

Scialom, T., Lamprier, S., Piwowarski, B., and Staiano, J. (2019). Answers unite! unsupervised metrics for reinforced summarization models. *arXiv preprint arXiv:1909.01610*.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Sellam, T., Das, D., and Parikh, A. P. (2020). Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Seng, D. and Wu, X. (2023). Enhancing the generalization for text classification through fusion of backward features. *Sensors*, 23(3):1287.

Shah, J. J., Smith, S. M., and Vargas-Hernandez, N. (2003). Metrics for measuring ideation effectiveness. *Design studies*, 24(2):111–134.

Shang, G., Ding, W., Zhang, Z., Tixier, A. J.-P., Meladianos, P., Vazirgiannis, M., and Lorré, J.-P. (2018). Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. *arXiv preprint arXiv:1805.05271*.

Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y. K., Saumya, S., and Roy, P. K. (2017). Predicting the "helpfulness" of online consumer reviews. *Journal of Business Research*, 70:346–355.

Singh, S. P., Kumar, A., Mangal, A., and Singhal, S. (2016). Bilingual automatic text summarization using unsupervised deep learning. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 1195–1200. IEEE.

Sjöbergh, J. (2007). Older versions of the rougeeval summarization evaluation system were easier to fool. *Information Processing & Management*, 43(6):1500–1505.

Sohn, K., Lee, H., and Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.

Song, W., Choi, L. C., Park, S. C., and Ding, X. F. (2011). Fuzzy evolutionary optimization modeling and its applications to unsupervised categorization and extractive summarization. *Expert Systems with Applications*, 38(8):9112–9121.

Srivastava, A. and Sutton, C. (2017). Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Steen, J. and Markert, K. (2021). How to evaluate a summarizer: Study design and statistical analysis for manual linguistic quality evaluation. *arXiv preprint arXiv:2101.11298*.

Steinberger, J. and Jezek, K. (2004). Using latent semantic analysis in text summarization and summary evaluation. *Proc. ISIM*, 4:93–100.

Suhara, Y., Wang, X., Angelidis, S., and Tan, W.-C. (2020a). Opiniondigest: A simple framework for opinion summarization. *arXiv preprint arXiv:2005.01901*.

Suhara, Y., Wang, X., Angelidis, S., and Tan, W.-C. (2020b). OpinionDigest: A simple framework for opinion summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5789–5798, Online. Association for Computational Linguistics.

Takamura, H. and Okumura, M. (2009). Text summarization model based on maximum coverage problem and its variant. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 781–789.

Tampe, I., Mendoza, M., and Milios, E. (2022). Neural abstractive unsupervised summarization of online news discussions. In *Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys) Volume 2*, pages 822–841. Springer.

Tang, H., Mi, Y., Xue, F., and Cao, Y. (2021). Graph domain adversarial transfer network for cross-domain sentiment classification. *IEEE Access*, 9:33051–33060.

Thakkar, K. S., Dharaskar, R. V., and Chandak, M. (2010). Graph-based algorithms for text summarization. In *2010 3rd International Conference on Emerging Trends in Engineering and Technology*, pages 516–519. IEEE.

Tohalino, J. V. and Amancio, D. R. (2018). Extractive multi-document summarization using multilayer networks. *Physica A: Statistical Mechanics and its Applications*, 503:526–539.

Toutanova, K., Brockett, C., Gamon, M., Jagarlamudi, J., Suzuki, H., and Vanderwende, L. (2007). The pythy summarization system: Microsoft research at duc 2007. In *Proc. of DUC*, volume 2007.

Tsarev, D., Petrovskiy, M., and Mashechkin, I. (2011). Using nmf-based text summarization to improve supervised and unsupervised classification. In *2011 11th International Conference on Hybrid Intelligent Systems (HIS)*, pages 185–189. IEEE.

Tsytsarau, M. and Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24:478–514.

Uzêda, V. R., Pardo, T. A. S., and Nunes, M. D. G. V. (2010). A comprehensive comparative evaluation of rst-based summarization methods. *ACM Transactions on Speech and Language Processing (TSLP)*, 6(4):1–20.

Vanderwende, L., Suzuki, H., Brockett, C., and Nenkova, A. (2007). Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 43(6):1606–1618.

Vanetik, N., Litvak, M., Churkin, E., and Last, M. (2020). An unsupervised constrained optimization approach to compressive summarization. *Information Sciences*, 509:22–35.

Voorhees, E. M. and Tice, D. M. (2000). Building a question answering test collection. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 200–207.

Wan, X. and Xiao, J. (2009). Graph-based multi-modality learning for topic-focused multi-document summarization. In *Twenty-First International Joint Conference on Artificial Intelligence*. Citeseer.

Wang, D., Zhu, S., Li, T., and Gong, Y. (2009). Multi-document summarization using sentence-based topic models. In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 297–300.

Wang, L., Shen, Y., Peng, S., Zhang, S., Xiao, X., Liu, H., Tang, H., Chen, Y., Wu, H., and Wang, H. (2022). A fine-grained interpretability evaluation benchmark for neural nlp. *arXiv preprint arXiv:2205.11097*.

Wei, F., Li, W., Lu, Q., and He, Y. (2008). Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 283–290.

Wei, F., Li, W., Lu, Q., and He, Y. (2010). A document-sensitive graph model for multi-document summarization. *Knowledge and information systems*, 22(2):245–259.

West, P., Holtzman, A., Buys, J., and Choi, Y. (2019). BottleSum: Unsupervised and self-supervised sentence summarization using the information bottleneck principle. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3752–3761, Hong Kong, China. Association for Computational Linguistics.

Whetten, D. A. and Cameron, K. S. (2005). Developing management skills. *(No Title)*.

Wilson, T., Wiebe, J., and Hoffmann, P. (2009). Articles: Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.

Wittrock, M. C. and Alesandrini, K. (1990). Generation of summaries and analogies and analytic and holistic abilities. *American Educational Research Journal*, 27(3):489–502.

Wu, H., Ma, T., Wu, L., Manyumwa, T., and Ji, S. (2020). Unsupervised reference-free summary quality evaluation via contrastive learning. *arXiv preprint arXiv:2010.01781*.

Xiao, W., Beltagy, I., Carenini, G., and Cohan, A. (2021). Primera: Pyramid-based masked sentence pre-training for multi-document summarization. *arXiv preprint arXiv:2110.08499*.

Xiao, Y., Zhao, T., and Wang, W. Y. (2018). Dirichlet variational autoencoder for text modeling. *arXiv preprint arXiv:1811.00135*.

Xie, S. and Liu, Y. (2008). Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4985–4988. IEEE.

Yao, J.-g., Wan, X., and Xiao, J. (2015). Compressive document summarization via sparse optimization. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Yih, W.-t., Goodman, J., Vanderwende, L., and Suzuki, H. (2007). Multi-document summarization by maximizing informative content-words. In *IJCAI*, volume 7, pages 1776–1782.

Yin, W. and Pei, Y. (2015). Optimizing sentence modeling and selection for document summarization. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Yogatama, D., Liu, F., and Smith, N. A. (2015). Extractive summarization by maximizing semantic volume. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1961–1966.

Yousefi-Azar, M. and Hamey, L. (2017). Text summarization using unsupervised deep learning. *Expert Systems with Applications*, 68:93–105.

Zhai, C., Cohen, W. W., and Lafferty, J. (2015). Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *ACM SIGIR Forum*, volume 49, pages 2–9. ACM New York, NY, USA.

Zhang, B., Li, H., Liu, Y., Ji, L., Xi, W., Fan, W., Chen, Z., and Ma, W.-Y. (2005). Improving web search results using affinity graph. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 504–511.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Zhang, M., Zhou, G., Huang, N., He, P., Yu, W., and Liu, W. (2023a). Asu-osum: Aspect-augmented unsupervised opinion summarization. *Information Processing & Management*, 60(1):103138.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, X., Zhang, R., Zaheer, M., and Ahmed, A. (2021). Unsupervised abstractive dialogue summarization for tete-a-tetes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14489–14497.

Zhang, Y., Callan, J., and Minka, T. (2002). Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 81–88.

Zhang, Z., Liang, X., Zuo, Y., and Li, Z. (2023b). Unsupervised abstractive summarization via sentence rewriting. *Computer Speech & Language*, 78:101467.

Zhao, B., Li, H., Lu, X., and Li, X. (2021). Reconstructive sequence-graph network for video summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2793–2801.

Zhao, B., Li, X., and Lu, X. (2018). Hsa-rnn: Hierarchical structure-adaptive rnn for video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7405–7414.

Zheng, H. and Lapata, M. (2019). Sentence centrality revisited for unsupervised summarization. *arXiv preprint arXiv:1906.03508*.

Zhong, S.-h., Liu, Y., Li, B., and Long, J. (2015). Query-oriented unsupervised multi-document summarization via deep learning model. *Expert systems with applications*, 42(21):8146–8155.

Zhu, W., Lu, J., Li, J., and Zhou, J. (2020). Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, 30:948–962.

Zhu, X., Goldberg, A. B., Van Gael, J., and Andrzejewski, D. (2007). Improving diversity in ranking using absorbing random walks. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 97–104.

Zopf, M., Mencía, E. L., and Fürnkranz, J. (2016). Beyond centrality and structural features: Learning information importance for text summarization. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 84–94.