Les Cahiers du GERAD ISSN: 0711–2440

An history of relevance in unsupervised summarization

F. Carichon, G. Caporossi

G–2023–53

November 2023

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2023
– Bibliothèque et Archives Canada, 2023

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2023
– Library and Archives Canada, 2023

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

An history of relevance in unsupervised summarization

Florian Carichon [a]

Gilles Caporossi [a, b]

[a] *Decision science Department,HEC Montréal, Montréal (Qc), Canada, H3T 2A7*

[b] *GERAD, Montréal (Qc), Canada, H3T 1J4*

florian.carichon@hec.ca
gilles.caporossi@hec.ca

November 2023
Les Cahiers du GERAD
G–2023–53

Abstract : Automatic document summarization aims at creating a shorter version of one or more documents to help users digest large amounts of information more easily by highlighting the most relevant material. Unsupervised methods are among the most suitable for this task since they do not require prior human intervention for condensing information. Therefore, this article provides a detailed analysis of the progress of unsupervised methods applied to document summarization, offering a better comprehension of the underlying fundamental principles that drive these approaches. With this goal in mind, this review addresses several important aspects. First it gives an overview of the field, with the related concepts, methods, and scenarios that allow an understanding of their context of application, their evaluation and their evolution over time. It also provides a new typology of analysis, clarifying the link between the task of document summarization and the definition of relevance, as seen in information theory, and how it influences the very construction of various systems. In-depth reflections are made on the relationship between certain contextual factors such as purpose and audience, and how they not only affect relevance, and thus the way important material is selected, but also ultimately influence the summarization task and its evaluation. Finally, we provide recommendations and research directions for integrating these insights into the field of automatic document summarization.

Keywords : Natural Language Processing, automatic document summarization, relevance, novelty, literature review

Résumé : Le résumé automatique de document a pour but de créer une version réduite d'un ensemble de textes pour aider des utilisateurs à mieux assimiler l'information pertinente contenue dans ces derniers. Les méthodes non supervisées sont parmi les méthodes les plus appropriées pour effectuer cette tâche puisqu'elles ne nécessitent pas d'étiquetage humain a priori pour condenser l'information. Cet article propose une revue de littérature détaillée de ces approches appliquées au résumé de document, octroyant alors une meilleure compréhension de leurs mécanismes et des principes fondamentaux sous-jacents. Cette analyse traite donc de plusieurs aspects importants. Tout d'abord elle offre une vue globale du domaine permettant d'expliquer certains contextes d'application, l'évaluation et l'évolution des approches non supervisées. Elle fournit une nouvelle typologie de classification des méthodes, clarifiant ainsi le lien entre la tâche du résumé de document et certains facteurs contextuels tels que l'intention et l'audience. Cela met particulièrement en avant comment ces facteurs influencent la notion de pertinence et la manière dont l'information est sélectionnée par un algorithme pour non seulement former le résumé, mais aussi mesurer sa performance. Résultant de ce travail, une discussion sur les ressemblances profondes entre les différentes méthodes et les limites qui leur sont reliées nous permet de proposer différentes recommandations et des pistes de recherche future afin d'intégrer les observations que nous avons réalisées.

Mots clés : Traitement du langage naturel, résumé automatique de documents, pertinence, nouveauté, revue de littérature

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 particularly 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 longer 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 finally email communications are all part of the phenomenon of information digitization. The volume and variety of text data available on the internet becomes unmatched by any other source. Moreover, its complementarity with more traditional structured data makes it an extremely valuable medium for companies and their analysts to understand better their economic environment, their consumers, and improve their decision-making [82]. In the finance and banking domain, we can take the example of the Net Promoter Score, which is one of the most used indicators to appreciate a customer's experience and compare institution efficiency [191]. 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 customers' voice 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 given rise to 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 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 [100, 149]. This process then encompasses a collection of different tasks that encounter these needs. Single long document, multi-documents, opinion oriented, aspect-based, or even update summarization are some well-known examples of such diversity. The first models were introduced in the late '50s and '60s and aimed to create abstracts of scientific articles in chemistry specifically [62, 145]. These first models were predicated on a set of heuristics combining statistical and linguistic methods to extract the information considered relevant. 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 very useful to provide 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 emergence 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 [13, 47, 118]. Obviously, many machine learning techniques can be employed and we refer the lecturer to the multiple literature review on these approaches [90, 105, 142]. Following the success of deep learning systems, several methods were also developed for both extractive [113] or abstractive summarization [198]. Finally, the breakthrough of using pre-trained Large Language Models (LLMs) based on transformers architectures such as

[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/
[4] https://www-nlpir.nist.gov/related_projects/tipster_summac/
[5] https://duc.nist.gov/
[6] https://tac.nist.gov/about/index.html

BERT [55], or GPT-2 and GPT-3 [188], or T5 [189] have recently allowed to obtain more meaningful representation with prior knowledge and methods as BART [128] or PEGASUS [240] are now the state-of-the art tools for abstractive summarization.

Whatever the method employed, it is based on a theoretical background coming from the behavioral study of humans when they must perform such a task. Authors in [97] 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 [97]. The way of approaching summarization copy this mechanism and is decomposed in three major steps as detailed in [164] in their survey:

1. Comprehension: Systems need to learn a representation of the original texts to fulfill users need. The representation will imply for example to focus on text representativeness with a graph-based strategy such as in *TextRank* [159], or to stress sentence specificity using TFIDF bag of words vectors [125]
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 [184]. They can further support diversity in sentence scoring with maximal marginal relevance approaches [27, 32]
3. Condensation: Systems must tackle the summary generation process. Optimization methods can impose sentence relevance while constraining summary length [157], and enforcing some linguistic features [77], or directly maximizing token likelihood [198]

These three steps, which make it possible to characterize all document summary approaches generally, they do not make it possible to distinguish and understand why certain models will perform better than others in certain contexts. Indeed, obviously a system producing a general-purpose summary, even a very good one, cannot respond to all tasks and all user needs [110]. 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 [100]:

- 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* [175].
  - 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 article 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, we can note that the representation of texts depends on input factors: a system based on the structure of the discourse to embody a text [152] cannot be applied to characterize multiple short tweets. The evaluation function will be affected by the user's needs. Textrank [159] depicts the central topic to inform about the subject mentioned while a model founded on topic modeling [87] 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 in relevant source texts such textual features as topic sentences, keywords, and repeated ideas [97]. It is therefore normal to notice the same phenomenon appearing in the summaries used by the automatic text summarization community, especially for supervised learning. Indeed, in this case the definition of relevance of the information and purpose become underlying since the operation of the algorithm and the characterization of the text and the importance of the information is done through the labels provided to the model. However, when no instructions are specified, the human summaries used, although different, are based on common properties such as the use of term frequency [166], including named entities, subject-specific terms, and the non-inclusion of reported facts and figures [86]. 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 [174] 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 [74]. This phenomenon thus establishes an implicit homogenization of summaries format toward specific indicative texts [108, 109]. 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 [222] depending on the tacit perception of the purpose of the abstract by the annotators[208]. 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 [160], 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 [204], or that they are easy to fool and that one can obtain very good 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 [208].

Whether for reasons of design flexibility, access to data, portability, or difficulty of evaluation, unsupervised methods have always been favored for document summarization and still are today with

the advent of deep learning and LLMs [114]. 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 labeling seems contradictory. The advantage of unsupervised models is therefore to 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. This 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 [90, 105, 142]. However, they remain superficial factors which may lack clarity to convey the underlying phenomena that could explain some 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 but also relate them to their personal knowledge base and experience [229]. Recognizing that, in most of the current summarization cases the target is almost always another person, the relation with the recipient's experience and knowledge should affect the generation process [98]. These individual attachments thus establish different abstract and unconscious intentions and purposes in the construction of the summary, impacting the perception of what significant information is [24]. 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 [123]. 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 deeply affect the functioning of the model and its performance [180]. Moreover, as it has been presented in [112], lot of 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 in articles, it is unfortunately never explicitly defined for most document summary approaches. This 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 analyzes of these said models and it limits the understanding of 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.

With the emergence of generative models and the new capabilities of LLMs, it seems essential to propose a literature review that not only provides an overview of historical and cutting edge systems, but also links the latter in the context of relevance and purpose factors pertained to automatic document summarization. Therefore, our contributions are the following:

- An exhaustive literature review covering unsupervised text summarization, from the historical approaches to the latest developments.
- A complimentary exploration of the dataset and evaluation metrics dedicated to unsupervised summarization.
- A new typology linking all these models, and based on the relationship between information relevance and its implications for text modeling, the formulation of the information selection function, and the final summary generation process. This review is the first to be particularly concerned in decomposing relevance into its topical and novelty constituents and in studying how the different elements characterizing these components relate to common attributes of unsupervised summarization methods and their evaluation.
- An analysis and discussion about the reasons behind these different notions of relevance observed. We take special interest in the impacts of the purpose factors defining summaries. We emphasize the link between the dimensions of indicativity and informativeness to topical relevance and between specificity and genericity to novelty relevance.

This article is structured as follows: The section 2 present the method used to collect and analyze the different papers referenced in this article. Section 3 presents the definition of relevance and how

the algorithms are classified relative to its various aspects. It also introduces the different data sets and evaluation metrics used for unsupervised summarization and the evolution of different community practices in that regard. Section 4 discusses the link between relevance and purpose factors and how it impacts the behavior of unsupervised algorithms and how it may bias its evaluation. Section 5 presents the limitation of our study as well as some future avenues of research. In section 6, we propose a general conclusion on our work.

Materials and methods

With the first works starting in the late 50s, automatic text summarization is one of the oldest and most explored fields of *Natural Language Processing* (NLP). Sentence compression, streams of information, applications to search engines, etc. in a supervised, semi-supervised, or unsupervised way, contribute to the richness of this domain. The objective of this section is thus to introduce which methods have been adopted in this article to understand how the collection of documents has been carried out, from which angle papers have been examined, and which criteria have been applied to include them in the results and analyses of this literature review.

Document summarization has a few conferences devoted to the discipline such as the Document Understanding Conference or the Text Analysis Conference, so we started by studying papers from these conferences. Automatic summarization is a discipline integrated to NLP and thus to the study of languages by means of a computer. The Association for Computational Linguistics (ACL) is the major organization dedicated to the field. Our data collection has therefore continued with the study of their publications and conferences. NLP being a specific topic of machine learning, we then broadened the scope of our research to include papers from journals of the discipline such as the Journal of Machine Learning Research (JMLR) or from the Association for the Advancement of Artificial Intelligence (AAAI), along with the study of the ArXiv database of preprints. We also integrated articles from significant multidisciplinary publishers such as Elsevier, Springer, IEEE, or CiteSeer into our study. Finally, we completed our collect with articles from Google Scholar searches, which now indexes a large majority of papers in this mainly open-source discipline. Once the sources were identified and selected, two distinct strategies were put in place to systematically collect articles addressing the topic of unsupervised methods in automatic text summarization. The first approach consisted in setting up, for each of the databases, a series of queries containing the following keywords: “unsupervised”, “non-supervised”, “weakly supervised” AND “text”, “document”, “content”, “sentence” AND “summarization”. Since the queries were applied to the title of the publications, these articles were automatically included in our data pool for study. However, one problem we have encountered is that many articles, both historical and recent, do not contain explicit statements that they are unsupervised in their title or abstract. Here are two examples to give the readers an idea of the difficulty to retrieve results through database queries in this context: “TextRank: Bringing Order into Texts” [160] or “Extractive Multi Document Summarization using Dynamical Measurements of Complex Networks” [216]. Consequently, to complete our data collection and to ensure that we had a complete view of the field, we decided to broaden the scope of the search to all documents containing the terms “text”, “document” AND “summarization” to ensure that we did not miss any major text in the field. However, to avoid adding too much noise to our study, we limited the dataset to articles with at least 100 citations. Finally, our second strategy consisted in examining numerous literature works dedicated to general document summarization. These literature reviews were retrieved via queries including the terms “advance”, “survey”, “state-of-the-art”, “review” AND “text”, “document” AND “summarization”. These queries allowed us to identify about twenty literature reviews on the domain, ranging from the early 2000s [153] to 2021 [63]. We collected all articles referenced in these works, thus allowing us to complete our list with papers considered foundational and recent articles in the discipline. The combination of these two strategies allows us to be confident that our data collection procedure included all major articles regarding text summarization. At the end of our process, just

over 400 articles were gathered for this literature review. We will now detail the methods, and criteria used to keep only relevant articles in our review on unsupervised automatic summarization.

As stated earlier, many articles do not directly mention the technical particularities or the specific scope of their summarization approach. A thorough reading of the methodological section, presenting the theoretical choices made to design the algorithms employed, was therefore carried out to understand where the system fits in. For each of these papers, we then looked to see if they met the criteria set for this analysis. Obviously, systems containing purely supervised methods were discarded from the study since we want to focus on unsupervised approaches. However, before going further in our selection process, it is important to recall that our study focuses not only on those methods but also on their relationship with the definition of relevance and novelty of information. Therefore, it was decided to remove papers dealing with tasks too specific because they could create a strong potential bias. We then did not consider the tasks of query-based, topic-oriented, guided, or update summarization in this article. Indeed, these tasks condition the way information is perceived since it is delineated before the summarization work. We did not also include papers addressing sentence compression because the input texts are very short and thus establish a bias on the redundancy of the information. Once these items were removed from the study, we also applied criteria directly concerning the unsupervised methods to preserve the focus of the research and reduce potential noise within the analysis. We therefore chose not to keep the unsupervised methods based on the creation of supervised features since the information retained upstream could be biased by a specific typology of training data. Finally, we minimized the number of articles that rely on minor modifications of more fundamental articles. As examples, we can mention the work on TextRank in the article [160] which is cited more than 4000 times and which is often simply adapted with a particular information representation or reapplied to a specific task or domain. It is essential to point out here that the manual analysis carried out during this selection phase includes a bias on the authors' interpretation of the context of use of the methods, on their supervised demeanor, or on their definition of the information relevance in specific settings. At the end of this selection process, our literature review includes 158 papers dedicated to unsupervised methods. These articles come mainly from the sources mentioned earlier in this section and cover a period from 1958 to the beginning of 2023. This selection of papers is completed by the addition of references concerning the data sets and the performance measures used in these articles, and by some theoretical work on manual document summarization and on information theory. This extension thus provides an exhaustive analysis in the results section as well as a comprehensive picture of the ecosystem surrounding unsupervised automatic text summarization. This also allows us to support our hypotheses and classification on the influence of relevance in designing unsupervised methods, and to propose a discussion on it impacts on the most popular and the state-of-the-art systems. Finally, we have included in this article, other literature review works on the domain to allow us locating our work in the literature, how to link it with previous works, and to understand its major contributions.

Finally, we would like to depict the procedure deployed to analyze the different methods presented. Our literature review is based on an inductive research approach. Indeed, we have noticed by reading numerous articles in the literature that the study and comparison of models are laborious and almost impossible. The main known explanation for this difficulty concerns the diversity of the data sets used and the baselines of references [53, 163]. In trying to understand why this variance existed between articles, we then established that several methods were employed in a context that did not correspond to the original purpose intended by its authors. While unsupervised methods highly rely on factors such as the data set, the target users, the task at hand, or the summary to be created, we have then observed that many authors do not explicitly specify these contextual factors behind the conception of their system. We therefore hypothesized that some of these contextual factors pertained to document summarization strongly influence on the notion of relevance, and that this relevance influenced the functioning of summarization models. To verify our hypothesis, we proceeded by conducting a survey of common practices used in unsupervised document summarization. We have then adopted a purely descriptive approach, reporting, and comparing the statements and methods presented in the selected articles to provide an overview of how state-of-the-art relates to our hypothesis. Therefore, we started

an exhaustive analysis of each article included in our work. More specifically, we reviewed the introduction and the related work sections to understand how the authors position their approach with others, and the task they intended to achieve. Then, an in-depth study of the methodology was performed to distinguish the various ways of representing and scoring the relevant information. The experimental processes and results were also examined to grasp the context in which the system was applied and how it compared to the rest of the literature. Once again, it seems essential to specify that these analyses suffer from an interpretation bias on our part concerning the statements and methods presented in these different articles. This bias is minimized by our work investigating information theories and crossing explanation from multiple recognized sources in the field. Using an empirical approach, we collected our results in a table allowing us to break down each method into its corresponding category, to indicate on which data set it was tested, and with which model it was compared. The results of this empirical work will be detailed on the results section of this article. Finally, although this typology does not solve all the problems related to the analysis of document summarization methods, it links this discipline to the larger field of information theory and natural language processing, thus allowing the discussion of fundamental phenomena that have been little explored in the rest of the literature.

1 Information content in unsupervised text summarization

Before going further in the analyses provided in the review, it seems important to provide some context about the fundamental task of summarization and why distinctions exist in the first place. A summary is defined as short statements that abridge the information and reflect the gist of the discourse of an original document [97]. The authors also explain that four 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 task, become a choice about what is the most important information in the source document [97]. To complete these tasks the writer of the summary must create relations between segments in the text but also relate them to their personal knowledge base and experience [229]. Recognizing that, in most of current summarization cases the target is almost always another person, the relation with the recipient's experience and knowledge should affect the generation process [98]. These individual attachments thus establish different abstract and unconscious intentions and purposes in the construction of the summary, impacting the perception of what is significant information. This is called an abnormal state of knowledge [24] and it influences the interpretation of important material in the documents. Automatic text summarization systems that can fulfill this task thereby need to respect exactly the same fundamental conditions of operation, especially in terms of identifying and selecting information. In the particular case of unsupervised models, the underlying structure of the data is used to generate an intermediate representation that links the input document and the produced summary. This intermediate structure is depicted by two different states that characterize the notion of information relevance: topical relevance, which identifies the information to be presented to the user; and novelty relevance, which takes into account the contribution of knowledge made to the user [123]. These general notions of information delineation will therefore also be influenced by the user's induced state of knowledge and their intention. In the remainder of this section, the objective is to define the notions of topical and novelty relevance, to characterize their observable distinctions, and to classify unsupervised summarization methods according to these notions.

1.1 Topical relevance

Since the work in [145], unsupervised models heavily depend on the notion of relevance to design the function that will evaluate importance scores to textual units. The definition of relevance has been extended and generalized as the measure that minimizes information loss between the text and its approximation [180]. 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 very 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* [93, 141] . While notions of *centrality* with the use of features similarity will be found [64, 185] 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.

1.1.1 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 [16]. Consequently, they capture key characteristics or events related to what they perceive as the major content in the original text [154]. 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 [18, 106]. The goal is then to assess the relevance of those terms to identify and select the one to preserve [211]. 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 [166]. Following this principle, multiple methods arise to score the importance of such information. The first techniques observe the assumptions of [145], where the word frequency characterizes importance. Further legitimized by the work of Nenkova et al. (2005) in their article [166], 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 [84, 85], conceptual units representing events [74, 213], or general keyphrases [194, 195]. 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 [76]. 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 articles use this metric to determine important terms [7, 41, 157, 167, 169]. 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 [34, 120]. In a similar vein, the authors in [73] 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 [79], information gain, and residual inverse document frequency are also used to score importance [142]. 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 [148], mutual information between the original sentence and the summary [228], or weighted-pooling operation on attention weights paid to keywords [241]. Finally, some new authors have characterized term importance directly using ROUGE score [132] 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 [29]. This framework displayed strong performances to improve the utilization of large language models to generate summaries in a zero-shot learning context [66, 197]. Another frequency-based hypothesis for isolating the contribution of terms in a unigram or multimodal language model [166, 220]. Specifically, we measure how probable a term

t is, given a trained probability distribution model on a background corpus [46, 249]. 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 [75]. 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) [33, 78, 206].

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 [62], but relative position scoring has been privileged later for its better performance [72, 171], and especially the average position of terms in the text [234]. 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 [72, 171]. 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 [62]. Some later researchers further specialized these cue expressions to specific fields, for example Arab politics [2]. Finally, some authors have used general knowledge bases such as Wikipedia to enrich the semantic information provided by some terms [200]. 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 [62], the use of numbers or dates [171, 201] or named entity [231] for newspapers, citations of other researchers for scientific papers [1], sentiment polarity and subjectivity for opinions [12], or even similarity to a query in query-focused summarization [107]. 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 [52]. Other authors also observed that traditional methods did not work optimally when summarizing opinions [214]. 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, authors in [8] 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* [133]. 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 [19]. Others propose rather to generalize the concepts to themes affirming that topics are related to each other [92]. 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 [87]. 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 [87]. *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 [218]. Furthermore, vectors obtained with NMF are sparser than with SVD, granting a better association between topics and sentences [126]. *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 [14, 224], or by a creating bipartite graph structure with topics and sentences and the use of importance calculation algorithms such as PageRank or HITS [178]. As for deep learning, some authors have shown that the application of autoencoders makes it possible to create concept-oriented vectors [237]. 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 [207]. The use of specific deep learning methods such as denoising techniques also allows isolating some concepts that are considered more important topical aspects [9]. 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 [89]. 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 [18]. 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 [18], but so can the relations between the members of the chain [30, 60], or their relative frequencies [89]. 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 [146]. A coherent text follows a specific discourse structure, and *Rhetorical Structure Trees* (RST) are objects meant to describe these relations between segments of text [172]. 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 [172], the nature of their connections [170], or promotion sets to characterize relations between nuclei and satellites [152]. Other improvements have been proposed over time and we refer the reader to the dedicated review of these models made by [219]. Other graph models close to RSTs, such as discourse graphs [40] or *Abstract Meaning Representation* (AMR) graphs [58], 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 [58] or domain-knowledge ranking systems [99] 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 [15]. 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 [26]. Finally, dependency trees have been used, especially because these structures show interesting properties for the abstractive generation of summaries [16, 37].

1.1.2 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 [184]. 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 [213] by being

the most redundant with other text segments [186], and that enforces a correlation of the semantic volume between the summary and the initial text(s) [236]. 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 [102].

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 [186]. We can then employ the text segments that are the closest to this centroid to form the summary [81, 209]. Multiple variations have been made to this method, either by changing the representation of the document(s) for richer ones such as word embeddings [122, 176, 196], by testing different clustering algorithms to use optimized fuzzy evolutionary algorithms [4, 209], as well as considering word-level features then grouping them to cluster interesting segments [17, 71], or hard singular value decomposition and maximizing proximity to the centroid through a greedy method to increase the semantic volume [236]. Some have considered adding semantic features such as WordNet information [102] or even full syntactic data, to evaluate paraphrasing and detect themes [20]. Once the cluster have been created, they can be used as input for abstractive methods by providing only central sentences to deep learning models [231]. The strategy in [45] proposes to create 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 [212]. Finally, authors in [10] 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 [64]. Therefore, some approaches have proposed to directly evaluate the closeness of each segment to every other [193]. 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 [64] because each link between vertices can be casting a vote or recommendation for those nodes [160]. Once the graph is constructed, there are many approaches to rank and select the most similar sentences. The LexRank algorithm [64] 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 [248], and the CoreRank system [69]. Authors in [160] 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 [215]. 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 [223]. 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 [182], special external resources and lexical features [96, 127], or structural metadata, by reinforcing links with intra-document information [226, 227], or with the hirearchical structure of the document [59]. Additional improvements create richer sentence representations with deep learning techniques to strengthen similarity estimation [3, 235, 246]. 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 [245]. Other authors have

proposed drastically different graph techniques to measure sentence centrality, such as using InfoMap and clustering coefficients [61], or by employing graph cuts to select subsets representing the summary then maximizing pairwise similarity through submodular optimization [113, 134, 136]. Finally, some researchers take a more direct approach by dynamically optimizing the pairwise similarity between passages of texts [157]. 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 [221]; or some have measured the coverage according to the capacity of a sentence to reconstruct the other sentences [138], 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 [4, 209], topic signatures [56], or by constituting homogeneous item sets of related sentences [193].

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 [233]. Others have used n-grams or embedding representations combined with Kullback-Leibler divergence to minimize the difference in the probability distribution of these elements [116, 181]. 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 [147]. 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 [140]. Thus, the authors have employed a paragraph vector model [124] 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 [207] or by using the new importance vectors in optimization processes [140, 247]. 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 [111]. 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 [203]. 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 [139, 203] 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 [39]. 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 [202], 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 [23]. 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 [42] 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 reproduc-

ing 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 [28]. 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 [243]. Finally, in for recent approaches, the average representation can also be used as an input for pretrained large language models [173].

1.1.3 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 in [211] and in [87]. 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 in [42] aims to represent a consensus between opinions, whereas the model proposed in [9] 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 [64] while salient information allows extracting key information that can be employed as an indicative summary [93]. 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.

1.2 Novelty relevance

In the first sections of this article, we have explored topical relevance in their work [6], 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 [180], 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 [123] 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 [238]. 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 [205]. 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 [43]. 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.

1.2.1 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 [143]. 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* [32]. 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 [32].

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 [201]. 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 [199, 216]. 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 articles. We can see the use of more traditional methods such as Jaccard or cosine similarities [78, 111, 216], or more evolved ones such as the use of mixture models [244], or combinations of Jensen and Kullback-Leibler divergences [217]. These similarity measures remain on lexical information and can perform less well than methods also relying on syntactic and semantic content [143]. 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 [94]. 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, authors in [183] 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 [141] 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 [18, 21, 26]. Interestingly, for recent abstractive approach, pairwise similarity is used as a preprocessing step to better guarantee the non-redundancy of the phrases explored [83].

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 in [32], 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 [137, 162], or to adapt the model to multi-document summarization [27]. 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 [39, 121]. Besides demonstrating better performance than strict similarity estimation [232], another benefit of this technique is that it creates summaries that are closely related to nonprofessional human summaries [192]. 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 [233], with methods such as pivoted-QR [46], or by project sentences on distant boundaries of a similarity graph [59] 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 [84, 157]. Some authors also proposed to enrich the redundancy estimation with various semantic features to perform better at the task of update summarization [161]. Similarly to this modification, other authors seeking optimal solutions have suggested submodular monotone functions [70] guaranteeing optimality with greedy algorithms, determinantal point process

algorithms [83], or nonmonotone graph-based functions with a high probability of optimality [134, 136]. 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 [115]. It has also been employed to manage the salience and the consistency of updated content for real-time streams of multiple tweets [129].

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 [239]. 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 [235]. 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 [5]. This approach has also demonstrated its ability to increase the coverage of various dimensions of an event in news summarization task [156]. 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 [166, 220]. 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 [235, 239]. 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 [248]; or by reinforcing previously explored ones, thereby decreasing the probability that the random walk ends on their neighbors [158]. Finally, another strategy used the structure of the networks to encourage random walks to visit external nodes dis

1.2.2 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 [158]. Indeed, by guaranteeing independence between selected segments, these approaches do not ensure access to diverse content [239]. 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 [238]. 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) [93] thus having a high probability of similar and overlapping content [1]. 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 [38]. The first methods select top-ranked segments of each cluster [1, 17, 91] 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 [186]. Other authors use complementary similarity measures such as cosine or Normalized Google distances [4, 209] 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. Authors in [39] employ 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, cluster opinionated features based on aspect detection are used to

ensure the coverage of different products aspect [179]. It is also possible to create clusters and used a trained language model to establish distinct consensual abstractive summaries for each aspect of customer reviews and thus diversify their presence in the output [28, 45]. 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 [212]. The model in [10] dynamically learns 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. Authors in [87] 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 [206] or bipartite word/topic graph [178] 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) [236]. 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 [236]. Several researchers have followed and modified this idea when applying deep learning techniques [31, 37]. 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 [51]. The approach deployed in [74] 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 [213]. Instead of considering conceptual units, we can employ bigrams and impose the constraint directly to have elements included in the summary only once [85]. Authors in [130], 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 [135] 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 [8]. 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 [173].

1.2.3 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 [59]. By minimizing the information repetition, we are trying to gain information, but we do not guarantee to cover a whole set of topics [239]. 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 [51, 249]. Whereas the same diversity-maximizing approaches perform extremely well when the objective is to cover various events in a news stream [156]. 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.

2 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.

2.1 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 [100, 110], 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 [53, 150]. 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 articles [50, 151, 174]. 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 [177], 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 [95], made up of roughly a million news stories and human-written abstractive summaries, which are related to a specific query; the Multi-News corpus [68] made for multi-document summarization and composed of more than 250 000 paired news and summaries; and finally the NewsRoom corpus [88] 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 [44, 239]. 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 [145]. It 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 [62], computer science [18], or medicine [178]. 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 [78], 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 [155], 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 [28, 42]. OPOSUM is another dataset based on Amazon products [11]. The dataset aims to provided human selected references that provides important aspects and extractive summaries focusing on slaience, popularity, fluency, and redundancy. [10] 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 [72]; IMDB movie reviews , which was in fact originally created for sentiment analysis [52]; or even manually designed ones. The last common kind of data used for automatic text summarization is meeting transcripts. The ISCI corpus [104] consists of 75 transcripts of naturally occurring meetings, where human annotators were asked to write 200-word abstractive and extractive summaries. The AMI Corpus [35] 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 [36, 48], emails [237], banking reports [58], or Wikipedia articles [2, 3].

For the 137 articles included in this review, Table 1 presents a distribution of the different corpora used in the evaluation of approaches. The different categories analyzed in these articles are news-based documents, opinion or blog data, scientific papers, meeting corpora, and others. If systems were applied on multiple corpora for evaluation, we count one occurrence for each category.

Table 1: Distribution of the data sets

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

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* [175]. 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.

2.2 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 [25, 67]. 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 [163]. 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 [54, 67, 230]. 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 discuss how they are related to topical and novelty relevance, the purpose factors, and how it influences the perception and the design of summarizers.

2.2.1 Intrinsic evaluation

Among the methods for evaluating intrinsic performance, there is a distinction between two categories for assessing the quality of a summary [150]. 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 [150]. 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) [131] used in the first years of the DUC conferences [49]. 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 [142]. 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. Article [187] 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 [165]. 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) introduces in his article [132] 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 articles, 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 [101], 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 [144]. 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) [163]. 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.

2.2.2 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 [151].

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 [150], 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 [100], 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* [65] and the work proposed in [204] 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 [165] 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 [168] 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 [150] 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 [100]. 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 [144]. However, these results do not show a good correlation with human productions. Other approaches such as *BertScore* [242] or *SUPERT* [80] have improved these results by measuring the *Word Mover's Distance* [119] 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 [150].

2.2.3 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 [142, 190] since they have not been applied in the analyzed articles 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 [67]. Once again, we perform a quantitative analysis of the different metrics used in the articles covered from this literature. The results of this analysis are displayed in Table 2 below.

Table 2: Distribution of the evaluation metrics

Metric	Number of articles
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

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 [67, 163, 210]. However, it is essential to note that most of the recent articles 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 [117]. 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 [163]. 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.

3 Discussion

Apart from a very recent article [114] 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 easily to various datasets, there is currently no systematic review dedicated to these methods and that aims to improve understanding of their fundamental disparities. 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 [114, 142]. In this work, we introduce a typology that takes its roots in the first analyses on the different dimensions characterizing a summary [100, 110]. While input and output factors are visible and their influence is easily observable on the functioning of models, purpose factors are more difficult to consider, as they create an implicit link between input and summary. However, the choice of a summary usage and target audience helps us to understand why the same segment of text may or may not be perceived relevant. This notion is rooted in the very foundations of information theory and the definitions of topical and novelty relevance, which identifies the information to be presented and the contribution of knowledge it will make to the user [123]. 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 [112, 180].

Indeed, this confirms our initial intuitions about the link between certain aspects of document summarization and information theory. More specifically, the contextual purpose factors are elements that connect the information contained in the input document(s) to that included in the output [110]. It's only natural, then, to see these factors intimately tied to the very definition and perception of relevance [180]. Two main characteristics materialize from contextual factors: the usage and the audience of the summary. When it comes to 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 the extent to which the semantic content meets the user's need for information and complements their previous knowledge. A clear relationship can then be seen between 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 distinct dimensions. Indicativity, which aims to promote content that enables understanding a topic in detail, and informativeness, which seeks to describe what is being said and the overall content. When we analyze the methods, we observe a first category that expresses the specific information of a document, spotlighting the characteristic elements of the input 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, one interesting property of indicative and informative content that we can observe in our classification and linked to the work in [149]: 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 a way, 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 obvious. 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 user by filtering out useless information [149], and genericness tends to provide a complete and generalized view of the source material by covering as much of its information as possible [149]. 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. And 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. There is indeed complementary information covered by specific and generic summaries, and both factors represent a spectrum where they are not incompatible, and systems can apply a combination of both redundancy and diversity to increase the covered content so long as the length constraints are respected [38, 102, 111]. Our contribution to the literature then becomes obvious, as 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 tries to address this issue [149].

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. We are therefore able to observe that not all systems are suited to all needs. Since the analysis of NLP models relies on comparison, it is essential to choose the right approaches for such a comparison. Otherwise, by not considering this control factor, the analysis becomes easily open to criticism, since it's 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 point of focus specific to purpose and unsupervised methods. Naturally, humans produce better summaries after being trained to identify in relevant source texts such textual features as topic sentences, keywords, and repeated ideas [97]. It is therefore 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 [166], including named entities, topic-specific terms [51], and the noninclusion of reported facts and figures [86]. Although many different guidance and tasks have been proposed at various 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 [174]. 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 [74]. Of course, one could argue that these issues are not important 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 and especially the ROUGE evaluation method [132], which represents 70% of evaluation metrics used in the literature and which is still being used as the almost sole measure in recent works. Other main methods used are the F1-score or the Pyramid method [165]. These intrinsic metrics suffer from several flaws [67] such as some bias towards lexical similarity and do not account for fluency and readability [204], or such as the fact that they are easy to fool, and that one can obtain very good 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 [208]. The major issue related to the observations made in this article is that they all use reference summaries created by human experts, which will inevitably have an impact on the way they operate, with all the issues we have just raised concerning the way 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 [108, 109]. 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 [175]. Given this new challenge, we can legitimately put forward the need to clarify these dimensions of purpose factors and to propose performance measurement metrics that are 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 specifically to describe in detail how the relevant information was considered, so that the appropriate methods and evaluations can be used. What's more, this new practice will not only be useful for unsupervised methods, but also for 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 therefore closer to all issues coming from human summarization too. 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 [194]. 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 encountered especially often in real-world applications and industries where there is a lot of specialized data with no gold standards available.

4 Limitations and future research

The main limitation of this litterature review stems from the fact that all our theory is drawn from the observation of 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 carried out an empirical examination of existing articles, and found this link between purpose factors and the behaving of different approaches. However, as we have already mentioned, our typology comes from a personal assertations and interpretation of the articles. Unfortunately, due to the lack of transparency in most of the articles 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? This lack of strict evaluation protocol, does not allow us to exactly measure, for example, if methods based of selectivity and redundancy wiil perform better than the ones based central and diverse information for a scenario where we seek to provide an indicative and specific view of an event. Moreover, as most papers compare models that were intended for different purposes and audience, it is extremely difficult to draw a formal conclusion on the type of methods to use to best meet certain tasks or user needs. This is why, even if this literature review relays statements already 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. This analysis is all the more limited in the case of pre-trained LLMs, since the information bias resulting from pre-processing and the complexity of analyzing all the layers making up these models make interpretation ever more complex.

We therefore naturally propose to continue studying the applicability of information relevance characteristics within large language models. In future research, our approach would make it possible to introduce a 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 [22], how information can modify prompt learning and thus the behavior of models [57], and finally how to complement methods for interpreting and analyzing results obtained in unique conditions [103, 225]. In the more specific context of automatic document summarization, understanding and exploring the differences in the behavior of pre-trained large languages models would follow the recommendations made in [114] 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 that the concepts of intention, purpose and type of text information be explicitly defined. 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 a real theory on the link between information and document summarization, but it will further fulfill the needs put forward in [25] on having alternative ways of understanding summarization approaches.

5 Conclusion

The emergence of the Internet has involved a large-scale digitization of the 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 information that exists. The interest in automatic text summarization thus has consequently become increasingly important in research but also in business communities. It has also established new opportunities and applications for the new data provided (email, scientific papers, medicine, blogs, reviews, etc.), the recently possible tasks (update, sentiment based, or personalized summarization), and the recent objectives they try to fulfill (answering questions, text overview, critics, etc.). Therefore, summarization systems are better understood and have seen great improvements, especially 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. This literature review then explored the use of unsupervised methods as a better way to address the task of text summarization. It highlights the relation between information theory and users' needs according to characteristics that go beyond superficial textual features. It first provides a clear framework for understanding how information is selected in unsupervised models and how building 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 way of perceiving information relevance. With the rise of 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 eve more relevant to understand how models capture relevance to propose systems that will answer these new needs.

References

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

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

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

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

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

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

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

[20] 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.

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

[22] 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.

[23] 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.

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

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

[26] 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.

[27] 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.

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

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

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

[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.

[32] 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.

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

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

[35] 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.

[36] 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.

[37] 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.

[38] 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.

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

[40] 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.

[41] 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.

[42] 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.

[43] 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.

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

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

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

[51] 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.

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

[53] 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).

[54] 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.

[55] 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.

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

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

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

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

[60] 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.

[61] 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.

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

[63] El-Kassas, W. S., Salama, C. R., Rafea, A. A., and Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. Expert Systems with Applications, 165:113679.

[64] 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.

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

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

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

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

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

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

[84] 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.

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

[86] 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.

[87] 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.

[88] 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.

[89] 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.

[90] 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.

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

[92] 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.

[93] 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.

[94] 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.

[95] 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.

[96] 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.

[97] 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.

[98] 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.

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

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

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] 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.

[106] 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.

[107] 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.

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

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

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

[111] 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.

[112] 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.

[113] 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.

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

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

[116] 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.

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

[118] 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.

[119] 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.

[120] 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.

[121] 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.

[122] 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.

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

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

[125] 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.

[126] 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.

[127] 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.

[128] 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.

[129] 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.

[130] 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.

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

[132] 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.

[133] 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.

[134] 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.

[135] 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.

[136] 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.

[137] 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.

[138] 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.

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

[140] 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.

[141] 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.

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

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

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

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

[146] 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.

[147] 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.

[148] 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.

[149] Mani, I. (2001a). Automatic summarization, volume 3. John Benjamins Publishing.

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

[151] 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.

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

[153] Maybury, M. (1999). Advances in automatic text summarization. MIT press.

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

[155] 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.

[156] 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.

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

[158] 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.

[159] 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.

[160] 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.

[161] 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.

[162] 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.

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

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

[165] 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.

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

[167] 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.

[168] 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.

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

[170] 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.

[171] 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.

[172] 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.

[173] 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.

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

[175] 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.

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

[177] 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.

[178] 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.

[179] 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.

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

[181] 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.

[182] 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.

[183] 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.

[184] 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.

[185] 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.

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

[187] 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.

[188] 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.

[189] 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.

[190] 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.

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

[192] 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.

[193] 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.

[194] 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.

[195] 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.

[196] 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.

[197] 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.

[198] Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685.

[199] 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.

[200] Sankarasubramaniam, Y., Ramanathan, K., and Ghosh, S. (2014). Text summarization using wikipedia. Information Processing & Management, 50(3):443–461.

[201] Schiffman, B., Nenkova, A., and McKeown, K. (2002). Experiments in multidocument summarization. Proceedings of HLT, pages 52–58.

[202] Schumann, R. (2018). Unsupervised abstractive sentence summarization using length controlled variational autoencoder. arXiv preprint arXiv:1809.05233.

[203] 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.

[204] Scialom, T., Lamprier, S., Piwowarski, B., and Staiano, J. (2019). Answers unite! unsupervised metrics for reinforced summarization models. arXiv preprint arXiv:1909.01610.

[205] Shah, J. J., Smith, S. M., and Vargas-Hernandez, N. (2003). Metrics for measuring ideation effectiveness. Design studies, 24(2):111–134.

[206] 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.

[207] 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.

[208] Sjöbergh, J. (2007). Older versions of the rougeeval summarization evaluation system were easier to fool. Information Processing & Management, 43(6):1500–1505.

[209] 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.

[210] 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.

[211] Steinberger, J. and Jezek, K. (2004). Using latent semantic analysis in text summarization and summary evaluation. Proc. ISIM, 4:93–100.

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

[213] 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.

[214] 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.

[215] 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.

[216] 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.

[217] 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.

[218] 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.

[219] 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.

[220] 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.

[221] Vanetik, N., Litvak, M., Churkin, E., and Last, M. (2020). An unsupervised constrained optimization approach to compressive summarization. Information Sciences, 509:22–35.

[222] 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.

[223] 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.

[224] 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.

[225] 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.

[226] 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.

[227] 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.

[228] 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.

[229] 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.

[230] 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.

[231] 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.

[232] 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.

[233] Yao, J.-g., Wan, X., and Xiao, J. (2015). Compressive document summarization via sparse optimization. In Twenty-Fourth International Joint Conference on Artificial Intelligence.

[234] 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.

[235] Yin, W. and Pei, Y. (2015). Optimizing sentence modeling and selection for document summarization. In Twenty-Fourth International Joint Conference on Artificial Intelligence.

[236] 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.

[237] Yousefi-Azar, M. and Hamey, L. (2017). Text summarization using unsupervised deep learning. Expert Systems with Applications, 68:93–105.

[238] 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.

[239] 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.

[240] 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.

[241] 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.

[242] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

[243] 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.

[244] 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.

[245] Zhang, Z., Liang, X., Zuo, Y., and Li, Z. (2023b). Unsupervised abstractive summarization via sentence rewriting. Computer Speech & Language, 78:101467.

[246] Zheng, H. and Lapata, M. (2019). Sentence centrality revisited for unsupervised summarization. arXiv preprint arXiv:1906.03508.

[247] 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.

[248] 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.

[249] 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.