

# Arabic Text Summarization Challenges using Deep Learning Techniques: A Review

Adnan Souri<sup>\*1</sup>, Mohammed Al Achhab<sup>1</sup>, Badr Eddine El Mohajir<sup>1</sup>, Mohamed Naoum<sup>2</sup>, Outman El Hichami<sup>3</sup>, Abdelali Zbakh<sup>4</sup>

<sup>1</sup>NTTI Team, FS, Abdelmalek Essaadi University, Tetouan, Morocco

\* a.souri@uae.ac.ma

<sup>2</sup>ENSIAS, Mohammed V University in Rabat, Rabat, Morocco

<sup>3</sup>Applied Mathematics And Computer Sciences Team,

ENS, Abdelmalek Essaadi University, Tetouan, Morocco

<sup>4</sup>ER-MSI Team, ENCGT, Abdelmalek Essaadi University, Tetouan, Morocco

**Abstract**— Text summarization is a challenging field in Natural Language Processing due to language modelisation and used techniques to give concise summaries. Dealing with Arabic language does increase the challenge while taking into consideration the many features of the Arabic language, the lack of tools and resources for Arabic, and the Algorithms adaptation and modelisation. In this paper, we present several researches dealing with Arabic Text summarization applying different Algorithms on several Datasets. We then compare all these researches and we give a conclusion to guide researchers on their further work..

**Keywords**- Extractive summarization; Abstractive summarization; Hybrid Approaches; Reinforcement Learning; Summarization of Arabic News; AraVec; MultiBooked

## I. INTRODUCTION

Research in Arabic Natural Language Processing (ANLP) has gained significant attention in recent years. Arabic is a challenging language for Natural Language Processing (NLP) due to its rich morphology, complex grammar, and unique linguistic characteristics.

One of the most important research areas in ANLP is text summarization [1], [2], [3], [4] and [5].

Text summarization aims to generate concise summaries of longer texts. Arabic text summarization research focuses on developing algorithms and techniques to handle the unique characteristics of Arabic text, such as complex sentence structures and the presence of rich morphological features [6] and [7].

This paper gives an overview of several research on Arabic text summarization. Several techniques are presented such as neural network architectures, convolutional neural networks, abstractive and extractive text summarization, the use of Transformer-based models, clustering and sentence scoring techniques and more other methods.

We then present some of Arabic language features that have the major role in dealing with ANLP. Among these features we present the feature of script itself, the diglossic nature of the language, grammatical and conjugational features.

We present then text summarization techniques and challenges that researchers encounter due to language

morphology, lexical ambiguity, sentence structure or limited resources.

Finally, we present results and analysis of our research paper with a discussion of the findings and their implications. And we conclude by giving the study contribution and future directions.

## II. RELATED WORK

Research on Arabic text summarization has gained significant attention in recent years. Here are some notable studies and approaches in the field:

In paper [1], the authors conduct a systematic review of automatic Arabic text summarization techniques. They explore the existing literature on Arabic text summarization and provide a comprehensive overview of the different approaches, methodologies, and algorithms used in this field.

The study aims to identify the key techniques employed in Arabic text summarization, including extractive and abstractive approaches, as well as the evaluation metrics and benchmarks commonly used for assessing summarization quality. The authors analyze the strengths and weaknesses of each technique and highlight the challenges specific to Arabic text summarization.

By conducting a systematic review, paper [1] provides insights into the current state-of-the-art in Arabic text summarization. They discuss the advancements made in the field and identify research gaps that require further investigation. The study contributes to the understanding of Arabic text

summarization techniques and serves as a valuable resource for researchers and practitioners in the field.

Ref. [2] conducts a comprehensive review of Arabic text summarization techniques. The review covers a wide range of approaches and methodologies used in Arabic text summarization, including both extractive and abstractive methods. The authors discuss various aspects of Arabic text summarization, such as linguistic complexities, domain-specific summarization, evaluation metrics, and datasets used in Arabic summarization research. They analyze the strengths and weaknesses of different techniques and approaches, providing insights into the current state-of-the-art in Arabic text summarization.

By conducting a thorough review, the study in [2] helps researchers and practitioners gain a comprehensive understanding of Arabic text summarization, its challenges, and the advancements made in the field. The authors also identify research gaps and highlight future research directions, aiming to contribute to the further development of Arabic text summarization techniques.

In [3], the authors present a survey of automatic Arabic text summarization techniques. The survey covers various aspects of Arabic text summarization, including both extractive and abstractive methods, as well as the challenges and evaluation metrics used in Arabic summarization research. The authors review and analyze the existing literature on Arabic text summarization, discussing the different techniques, algorithms, and approaches employed in this domain. They explore the linguistic complexities of the Arabic language and how they impact the development of effective summarization models. They identify the strengths and limitations of different approaches and highlight the research gaps that require further exploration.

In [4], the authors present a comprehensive state-of-the-art survey of feature-based automatic text summarization methods. The survey focuses on summarization techniques that utilize various features, such as statistical, linguistic, semantic, and syntactic features, to extract salient information from text and generate summaries.

The authors in [4] review and analyze the existing literature on feature-based automatic text summarization, discussing the different approaches, algorithms, and evaluation methods employed in this field. They examine how different features contribute to the summarization process and highlight the strengths and limitations of various feature-based methods. They identify the trends, challenges, and potential research directions in this area.

Meanwhile, In [5], the authors present a comprehensive review of automatic text summarization methods. The review covers various techniques, algorithms, and approaches used in the field of automatic text summarization. The authors provide

an overview of the different methods and evaluate their strengths, limitations, and potential applications.

For the same, in [6], the authors present a comprehensive review of automatic text summarization techniques and methods. The review covers various approaches and algorithms employed in the field of automatic text summarization. The authors provide an overview of the different techniques, evaluate their effectiveness, and discuss their strengths, limitations, and potential applications.

In [7], authors research work focuses on multi-document summarization in Arabic using neural network architectures. The study proposes a model that leverages the Transformer-based architecture and utilizes attention mechanisms to generate summaries from multiple input documents.

This study presented in [8] explores the application of deep learning techniques, such as convolutional neural networks (CNN) and recurrent neural networks (RNN), for Arabic text summarization. It compares the performance of different deep learning architectures in generating abstractive summaries.

Reference [9] investigates abstractive Arabic text summarization using neural network models. The study explores the use of sequence-to-sequence models with attention mechanisms to generate concise summaries while capturing the semantic meaning of the input text.

In [10], authors investigate the use of Transformer-based models, specifically BERT (Bidirectional Encoder Representations from Transformers), for Arabic news summarization. The study explores different fine-tuning approaches and compares the performance with traditional methods.

The authors in [11] propose an extractive summarization approach for Arabic text using lexical chains and sentence similarity. Lexical chains are built to capture semantic relationships between sentences, and sentence similarity measures are used to identify important sentences for the summary.

In [12], authors propose an extractive summarization method specifically for Arabic news articles. They use sentence and phrase-level features, such as term frequency and position, to rank sentences and select the most relevant ones for the summary.

The study in [13] combines clustering and sentence scoring techniques for Arabic text summarization. Clustering is used to group similar sentences, and sentence scoring based on features like term frequency-inverse document frequency (TF-IDF) and sentence position is employed to select representative sentences from each cluster.

In [14], the research paper presents an approach that combines sentence compression and semantic similarity techniques for Arabic text summarization. Sentence compression is applied to reduce sentence length, and semantic

similarity is used to select the most representative sentences for the summary.

The work presented in [15] explores the application of abstractive summarization techniques to Arabic news articles. The proposed method incorporates deep learning models, such as Long Short-Term Memory (LSTM) networks, to generate summaries by paraphrasing and restructuring the original text.

In [16], the study proposes a sentence ranking approach based on connectedness and positional weight to generate extractive summaries in Arabic. The approach takes into account sentence cohesion and coherence to select important sentences.

This study in [17] introduces an extractive summarization approach for Arabic text based on sentence clustering. The method clusters sentences using a semantic similarity measure and selects representative sentences from each cluster to form the summary.

The research in [18] proposes a method for Arabic text summarization based on the concept of lexical chains. Lexical chains are used to identify and extract important concepts from the text, which are then used to generate summaries by selecting relevant sentences.

In [19], a survey provides an overview of the challenges, techniques, and evaluation methods in Arabic text summarization. It discusses extractive and abstractive approaches, linguistic features, and domain-specific summarization.

For the same, in [20], a survey provides an overview of different approaches and techniques used in Arabic text summarization, including extractive and abstractive methods. It discusses challenges specific to Arabic and highlights key research directions in the field.

These studies highlight the ongoing research efforts in Arabic text summarization, addressing the unique challenges of the Arabic language and exploring various techniques, ranging from traditional feature-based methods to state-of-the-art deep learning approaches. Researchers continue to explore novel approaches, develop specialized datasets, and refine evaluation metrics to advance the field of Arabic text summarization.

The Arabic language is a Semitic language and is widely spoken across the Middle East and North Africa. It has a rich history and is one of the six official languages of the United Nations. Here are some key features of the Arabic language:

Arabic uses a script known as the Arabic alphabet, which consists of 28 letters, some of which have one form, while others have two forms, three or four forms. The table 1 presents some of these letters showing the number of forms. Letters forms are depending on their position within a word. The script is written from right to left.

TABLE I. LETTERS FORMS ACCORDING TO THE POSITION IN A WORD

Number of forms	Forms	Letter
1	" د "	(dal) د
2	" س " ; " سين "	(seen) س
3	" ه " ; " هـ " ; " هاء "	(haa) هـ
4	" ع " ; " عـ " ; " عا " ; " عاء "	(ayn) ع

Arabic has a diglossic nature, meaning that it has two forms: Classical Arabic (also known as Quranic Arabic) and Modern Standard Arabic (MSA). Classical Arabic is the language of the Quran and is used in formal settings, literature, and religious texts. MSA is used in formal speech, writing, and media, while different dialects are used for everyday conversation.

Like other Semitic languages, Arabic is based on a root system. Words in Arabic are derived from a three-consonant root, and different forms and meanings are created by adding vowels, prefixes, and suffixes to these roots as mentioned in table 2.

TABLE II. DERIVATION BY ADDING VOWELS, PREFIXES AND SUFFIXES

English	Arabic
He wrote	(kataba) كتب
He writes	(yaktubu) يكتب
A book	(kitabun) كتاب
A library	(maktabatun) مكتبة
written	(maktubun) مكتوب
Children quran school	(kuttabun) كتاب

Arabic has grammatical gender, with nouns and adjectives being either masculine or feminine. Agreement between nouns, adjectives, and verbs is important in Arabic, meaning that they must agree in gender, number, and case.

Arabic has a dual form for nouns, pronouns, and verbs, which is used to indicate exactly two of something. Table 3 shows the case of a sentence in the singular form, dual and then the plural form.

TABLE III. AGREEMENT BETWEEN NOUNS AND ADJECTIVES

The sentence in English	The sentence in Arabic
The girl is beautiful	الفتاة جميلة
The two girls are beautiful	الفتاتان جميلتان
The girls are beautiful (three or more)	الفتيات جميلات

Arabic verbs are highly inflected and conjugated according to tense, mood, aspect, person, and number. There are several verb forms, known as "conjugations," which are used to indicate different aspects and time frames. In table 4, we present the verb "to write" conjugated in the present tense with some pronouns.



TABLE IV. A VIEW OF CONJUGATION OF VERB (KATABA) TO WRITE

English	Arabic
I write	(aktubu) اكتب
you write (masc.)	(taktubu) تكتب
you write (fem.)	(taktubiina) تكتبين
you write (dual)	(taktubaani) تكتبان
you write (masc.) (plural)	(taktubuuna) تكتبون
he writes	(yaktubu) يكتب

Arabic has separate subject and object pronouns, which are used in different positions within a sentence. Pronouns can also be attached as suffixes to verbs, prepositions, and nouns to indicate possession or object agreement.

Arabic has a relatively small vowel system, consisting of long and short vowels. Short vowels are often not written, but they are important for pronunciation and are indicated by diacritical marks in certain texts.

Arabic has a variety of sounds that may be unfamiliar to speakers of other languages. It includes guttural sounds like "ayn" (/ʕ/) and "ha" (/ħ/), which can be challenging for non-native speakers to pronounce.

Arabic calligraphy is a highly regarded art form, it is like fonts in text processing software but with an artistic aspect. The beauty of the Arabic script is often emphasized in various forms of artistic expression, such as Quranic manuscripts, architecture, and decorative arts.

Due to all these Arabic language features and more others, dealing with ANLP is quite challenging.

### III. ARABIC TEXT SUMMARIZATION TECHNIQUES

Arabic text summarization is the task of generating concise summaries of longer Arabic texts while preserving the key information and main points [1], [2], [3], [4] and [5]. It faces unique challenges due to the complex structure of Arabic sentences, rich morphology, and the presence of dialectal variations [6] and [7]. In this paragraph, we present some approaches and techniques used in Arabic text summarization:

#### A. Extractive Summerization

Extractive methods involve selecting important sentences or phrases from the original text to construct a summary [21]. Key techniques for extractive summarization in Arabic include:

- **Sentence Scoring:** Sentences are assigned scores based on various features such as word frequency, position, or semantic similarity to identify important content.
- **Graph-based Algorithms:** Text is represented as a graph, where nodes represent sentences, and edges represent relationships between them. Graph algorithms are used to identify the most informative sentences for the summary.

- **Clustering:** Similar sentences are grouped together, and representative sentences from each cluster are selected to form the summary.

#### B. Abstractive Summerization

Abstractive methods aim to generate summaries by paraphrasing and rephrasing the original text, rather than directly selecting sentences [22]. These methods involve more advanced natural language processing techniques, such as:

- **Sequence-to-Sequence Models :** Neural network architectures, such as Recurrent Neural Networks (RNNs) or Transformer models, are trained to generate summaries by learning to map input text to output summaries. These models can capture the semantic and contextual information necessary for generating abstractive summaries.
- **Attention Mechanisms :** Attention mechanisms allow the model to focus on different parts of the input text while generating the summary. They help in aligning important content and generating more coherent summaries.

#### C. Hybrid approaches

Some approaches combine extractive and abstractive techniques to leverage the advantages of both. These methods first extract important sentences or phrases and then employ abstractive methods to rephrase and reorganize the extracted content into a more concise and coherent summary.

#### D. Linguistic Features

Linguistic features specific to Arabic can be used to aid the summarization process. These features include part-of-speech (POS) tags, morphological analysis, and syntactic information, which can assist in identifying key phrases, relationships between sentences, and important content for the summary.

#### E. Supervised Learning

Some approaches use supervised learning algorithms like in [23], where models are trained on pairs of input texts and corresponding summaries. These models learn to generalize from the training data to generate summaries for new texts.

#### F. Evaluation

Various evaluation metrics are used to assess the quality of Arabic text summaries. Metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Evaluation Understudy) are commonly employed to measure the similarity between generated summaries and reference summaries or human judgments.

#### IV. ARABIC TEXT SUMMARIZATION CHALLENGES

Arabic text summarization faces several challenges due to the unique characteristics of the Arabic language and the complexities of summarizing Arabic text. Here are some key challenges in Arabic text summarization:

##### A. Rich morphology

Arabic has a rich morphological structure, with extensive use of prefixes, suffixes, and root-based word formation. This complexity adds challenges to the identification and extraction of important content for summarization.

##### B. Sentence structure

Arabic sentences can be long and have complex syntactic structures, including nested clauses and subclauses. Understanding and representing these structures accurately in the summary is a challenge.

##### C. Lexical ambiguity

Arabic words often have multiple meanings and interpretations, which can lead to ambiguity in the summarization process. Resolving lexical ambiguity accurately is crucial for generating meaningful and coherent summaries.

##### D. Limited resources

Compared to some other languages, there is a relatively limited availability of high-quality Arabic summarization datasets, linguistic resources, and tools. The scarcity of resources makes it challenging to train and evaluate robust summarization models.

##### E. Cross-domain Generalization

Arabic text summarization models trained on specific domains may struggle to generalize well to different domains. Adapting and fine-tuning models for specific domains and achieving cross-domain generalization is a challenge in Arabic text summarization.

##### F. Evaluation Metrics

The evaluation of Arabic text summarization can be challenging due to the lack of widely accepted evaluation metrics specifically tailored for Arabic summaries. Existing metrics, such as ROUGE, may not fully capture the linguistic nuances and intricacies of Arabic summaries.

##### G. Text complexity and topic variations

Arabic text covers a wide range of topics, including news, social media, legal documents, and scientific articles. Summarization systems must handle diverse topics and adapt to varying text complexity levels to produce accurate and informative summaries.

Addressing these challenges requires ongoing research and development of specialized techniques, datasets, and evaluation

methodologies specifically designed for Arabic text summarization. The exploration of advanced NLP methods and deep learning approaches can help to overcome these challenges and improve the quality and effectiveness of Arabic text summarization systems.

#### V. DATASETS

##### A. Summarization of Arabic News (SAN)

In [24], the authors address the need for a dataset tailored to Arabic single-document summarization. Single-document summarization focuses on generating concise and informative summaries from individual documents, which is an essential task in natural language processing and information retrieval.

The authors recognize that existing summarization datasets may not fully capture the linguistic and contextual peculiarities of the Arabic language. To address this gap, they propose the creation of a new dataset that aligns with the unique challenges and characteristics of Arabic text summarization [24].

To build the Arabic single-document summarization dataset, the authors describe the data collection and annotation process. They carefully curate a diverse set of Arabic documents from various domains and topics to ensure the dataset's representativeness.

Human annotators are then employed to create reference summaries for each document in the dataset. These reference summaries are crucial for evaluating the performance of summarization systems using standard evaluation metrics like ROUGE.

The authors discuss the importance of the dataset in advancing research in Arabic single-document summarization. The availability of a well-constructed dataset allows researchers to develop and evaluate state-of-the-art summarization models specifically tailored to the Arabic language.

In conclusion, in [24], the authors present the efforts to create an Arabic single-document summarization dataset to address the need for comprehensive and language-specific resources in the field. The dataset contributes to the development and evaluation of Arabic summarization systems, ultimately enhancing the accessibility and usability of automated summarization technology in Arabic text processing.

In [25], the authors address the task of Arabic single-document summarization, which aims to generate concise and coherent summaries from individual Arabic documents. The proposed approach utilizes lexical chains to extract salient information from the document and construct a coherent summary.

The authors begin by explaining the concept of lexical chains, which are sequences of related words or terms that share semantic relationships within a document. These chains help identify important concepts and topics present in the text.



The summarization process involves several steps: pre-processing the document to remove noise and irrelevant information, building lexical chains to identify key concepts, and selecting sentences from the document that best represent these concepts to form the summary [25].

To build lexical chains, the authors in [25] use linguistic features such as word co-occurrence and semantic similarity. By identifying chains of related terms, the method aims to capture the core ideas and topics covered in the document.

The selected sentences are then ranked based on their relevance to the identified concepts, and the top-ranked sentences form the final summary.

The authors evaluate the performance of their summarization approach using standard evaluation metrics and compare it with other summarization methods [25]. The results demonstrate the effectiveness of the proposed method in generating meaningful and coherent Arabic summaries.

In conclusion, the research in [25] presents a novel approach to Arabic single-document summarization based on lexical chains. By leveraging semantic relationships among words, the method identifies key concepts and constructs a concise summary that represents the main ideas in the document. The research contributes to the development of effective summarization techniques for the Arabic language, advancing the state-of-the-art in Arabic natural language processing.

The dataset can be accessed through the SAN GitHub repository at [26].

#### B. *AraVec*

In [27], the authors present AraVec, which is a set of Arabic word embedding models that have been trained on a large corpus of Arabic text. Word embeddings are vector representations of words in a high-dimensional space, where words with similar meanings or usage patterns are represented by vectors close to each other.

The authors explain that word embeddings play a crucial role in various natural language processing tasks, such as text classification, sentiment analysis, and machine translation. AraVec provides pre-trained word embeddings specifically tailored to the Arabic language, enabling researchers and practitioners to leverage these embeddings in Arabic NLP applications without the need for additional training data.

To create AraVec, the authors, in [27], likely used a large corpus of Arabic text, such as news articles, web content, or social media data. They would have employed popular word embedding techniques like Word2Vec, GloVe, or FastText to generate the word embeddings.

The availability of AraVec provides a valuable resource for the Arabic natural language processing community, allowing researchers and developers to easily incorporate high-quality Arabic word embeddings into their NLP projects. The dataset

can foster the development of more accurate and effective Arabic language models and applications.

In [28], the authors propose an abstractive approach to Arabic text summarization using deep learning techniques. Abstractive summarization involves generating concise and coherent summaries by paraphrasing and rephrasing the content of the source document, rather than extracting sentences directly.

The authors aim to address the challenges of Arabic text summarization, including the linguistic complexities and morphological richness of the Arabic language. Deep learning models are employed to learn the semantic representation and structure of the input text, enabling the generation of meaningful and grammatically correct summaries [28].

The proposed approach likely involves a sequence-to-sequence (seq2seq) architecture, commonly used for abstractive summarization tasks. The model takes an Arabic document as input and encodes it into a fixed-length vector representation. Then, a decoder generates the summary by predicting a sequence of words based on the encoded representation.

To train the deep learning model, the authors in [28] would require a dataset of paired Arabic documents and their corresponding human-generated summaries. Depending on the availability of appropriate datasets, they might use existing resources or create their own dataset.

The research in [28] presents a deep learning-based approach to abstractive Arabic text summarization, aiming to overcome the challenges posed by the Arabic language. The research contributes to the development of advanced summarization techniques for Arabic, with potential applications in information retrieval, natural language processing, and content summarization.

#### C. *MultiBooked*

MultiBooked dataset serves as a benchmark for multilingual multi-document summarization [29] and [30]. The dataset aims to provide a comprehensive and multidimensional resource for evaluating summarization models. It includes news articles collected from various Arabic news sources, covering a wide range of topics and genres [29].

The MultiBooked corpus contains a diverse collection of documents in multiple languages, making it suitable for multilingual summarization research. The data is curated from various sources and covers a wide range of topics and genres, ensuring that the corpus is representative of real-world scenarios [29] and [30].

The authors in [29] describe the process of creating the dataset, which involves collecting the source documents and obtaining multiple human-generated summaries for each document. The dataset is carefully annotated to ensure high-quality summaries that capture the essence of the source texts.

The resulting dataset serves as a benchmark for evaluating the performance of Arabic text summarization systems [29].

The research in [29] highlights the importance of the MultiBooked dataset in advancing research in Arabic text summarization. It provides researchers with a valuable resource for training and evaluating summarization models, particularly in the context of multi-document summarization. The authors discuss potential use cases and applications of the dataset and emphasize its potential impact on the development of Arabic summarization technologies.

Overall, the study by [29] showcases the MultiBooked dataset as a multidimensional resource that contributes to the advancement of Arabic text summarization research. Researchers can refer to this reference for more detailed information about the dataset and its usage in Arabic text summarization studies.

In [30], The authors explain that multi-document summarization is a challenging natural language processing task where the objective is to generate a concise summary that captures the essential information from a set of related documents. This type of summarization is particularly important in scenarios where information is scattered across multiple sources, such as news articles or research papers.

To create the dataset, human annotators generated summaries for each document cluster in the corpus. These summaries were carefully crafted to capture the main ideas and key information from the source documents. The presence of human-generated summaries allows researchers to evaluate the performance of their summarization models using standard metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [30].

The authors in [30] also discuss the importance of a standardized benchmark for evaluating multilingual multi-document summarization systems. The availability of such a benchmark fosters the development of robust and effective summarization algorithms that can operate across different languages and domains.

In conclusion, the MultiBooked corpus presented in this study provides a valuable resource for researchers working on multilingual multi-document summarization. It offers a diverse and representative set of documents in multiple languages, allowing researchers to advance the state-of-the-art in this challenging NLP task. The dataset has the potential to facilitate the development of practical summarization systems capable of handling information from various sources and domains.

The dataset can be accessed through the MultiBooked GitHub repository at [31].

## VI. ALGORITHMS

Arabic text summarization algorithms are designed to generate concise and coherent summaries from Arabic text.

These algorithms can be categorized into main types : Extractive Summarization, Abstractive Summarization, Hybrid Approaches, Rule-based Approaches, and Deep Reinforcement Learning. Here's an overview of each type and some commonly used algorithms:

### A. *Extractive Summarization Algorithms:*

Extractive summarization algorithms select and extract sentences or phrases directly from the source text to form the summary. These algorithms identify the most relevant and important sentences based on certain scoring criteria. Some popular extractive summarization algorithms include:

1) *Term Frequency-Inverse Document Frequency (TF-IDF)* : This algorithm ranks sentences based on the importance of the words they contain relative to the entire document. Sentences with higher TF-IDF scores are more likely to be included in the summary.

2) *TextRank*: Inspired by Google's PageRank algorithm, TextRank treats sentences as nodes in a graph and uses edge weights to represent the relationships between sentences. It ranks sentences based on their centrality in the graph, and the top-ranked sentences are selected for the summary.

3) *LexRank*: Similar to TextRank, LexRank uses sentence similarity as edge weights in the graph representation. It selects sentences that are both important and diverse to ensure a well-rounded summary.

4) *KL-Sum*: KL-Sum uses Kullback-Leibler divergence to measure the information loss when a sentence is removed from the document. Sentences with the least information loss are included in the summary.

5) *Supervised Learning with Features* : Traditional supervised learning algorithms, combined with linguistic features and domain-specific knowledge, have also been used for extractive summarization.

### B. *Abstractive Summarization Algorithms:*

Abstractive summarization algorithms generate summaries by paraphrasing and rephrasing the content of the source text. These algorithms create new sentences that convey the key information in a more concise form. Some commonly used abstractive summarization techniques include:

1) *Sequence-to-Sequence (seq2seq) models*: seq2seq models, often based on Recurrent Neural Networks (RNNs) or Transformer architecture, encode the source text into a fixed-length vector and then use a decoder to generate the summary.

2) *Pointer-Generator Networks*: These models combine extractive and abstractive techniques. They can copy words from the source text (extractive) while also generating new words (abstractive) to form the summary.

3) *c. BERT-based models*: Bidirectional Encoder Representations from Transformers (BERT) and other language



models can be fine-tuned for abstractive summarization, allowing them to understand context and generate summaries effectively.

#### C. Pointer-Generator Networks

This model combines extractive and abstractive approaches by using a pointer mechanism to select words directly from the input document when generating the summary.

#### D. Reinforcement Learning

Some research has explored using reinforcement learning to fine-tune summarization models by optimizing evaluation metrics like ROUGE

#### E. Latent Dirichlet Allocation (LDA)

LDA is a topic modeling algorithm that can be adapted for document summarization by selecting sentences that represent the main topics in the document.

It's important to note that Arabic text summarization algorithms may have specific adaptations to address the unique linguistic characteristics of the Arabic language, such as its rich morphology and right-to-left writing direction.

## VII. CONCLUSION

Arabic summarization, whether single-document or multi-document, faces several limitations due to the unique characteristics of the Arabic language and the available resources. As a conclusion of this paper, we present several reasons of these limitations.

Annotated datasets specifically designed for Arabic summarization are scarce compared to resources available for other languages like English. The lack of large-scale and diverse datasets hampers the development and evaluation of Arabic summarization models.

Arabic has complex linguistic features, including rich morphology, intricate syntax, and a wide range of dialects. Capturing these characteristics accurately in summarization models is challenging. Morphological analysis, dealing with dialectal variations, and maintaining fluency and coherence are ongoing research areas.

Arabic summarization for specific domains, such as legal, medical, or scientific texts, requires domain-specific resources and expertise. However, such resources are often limited in Arabic, hindering the development of domain-adaptive summarization models.

Compared to English for example, the availability of NLP tools and resources for Arabic is relatively limited. This includes robust part-of-speech taggers, syntactic parsers, named entity recognition systems, and sentiment analysis tools. The absence of these resources affects the performance and accuracy of Arabic summarization models.

Abstractive summarization, which generates novel summaries, is more difficult in Arabic due to its complex morphology and syntax. Ensuring semantic coherence, maintaining the original meaning, and avoiding grammatical errors in generated summaries present significant challenges.

Evaluating the quality of Arabic summaries is a challenge. Metrics like ROUGE are widely used, but they may not fully capture the nuances of the Arabic language. Developing reliable evaluation metrics and benchmarks specific to Arabic summarization remains an ongoing research area.

Training and fine-tuning large-scale summarization models require substantial computational resources. Access to high-performance computing infrastructure and large-scale pre-training data can be a limitation for researchers and practitioners, particularly in resource-constrained settings.

Researchers and developers are actively working to address these limitations by creating new datasets, improving language-specific tools and resources, and developing novel techniques tailored to the unique characteristics of the Arabic language. Continued efforts in these areas will help overcome the limitations and advance the field of Arabic summarization.

## REFERENCES

- [1] K. J. Abdelqader, A. Mohamed, and K. Shaalan, "Systematic Review of Automatic Arabic Text Summarization Techniques". In International conference on Variability of the Sun and sun-like stars: from asteroeismology to space weather (pp. 783-796). Springer, Singapore, 2023.
- [2] A. Elsaid, A. Mohammed, L.F. Ibrahim, and M.M. Sakre, "A comprehensive review of arabic text summarization". *IEEE Access*, 10, 38012-38030, 2022.
- [3] M. A. Elmenshawy, T. Hamza, and R. El-Deeb, "Automatic arabic text summarization (AATS): A survey". *Journal of Intelligent & Fuzzy Systems*, 43(5), 6077-6092, 2022.
- [4] D. Yadav, R. Katna, A.K. Yadav, and J. Morato, "Feature Based Automatic Text Summarization Methods: A Comprehensive State-of-the-Art Survey". *IEEE Access*, 10, 133981-134003, 2022.
- [5] D. Yadav, J. Desai, and A.K. Yadav, "Automatic text summarization methods: A comprehensive review". *arXiv preprint arXiv:2204.01849*, 2022.
- [6] A.P. Widyassari, S. Rustad, G. F. Shidik, E. Noersasongko, A. Syukur, and A. Affandy, "Review of automatic text summarization techniques & methods". *Journal of King Saud University-Computer and Information Sciences*, 34(4), 1029-1046, 2022.
- [7] Elaraby, and Mourad, "Multi-Document Arabic Summarization Using Neural Networks". 2021.
- [8] Elmahdy et al., "Arabic Text Summarization Based on Deep Learning Techniques", 2020.
- [9] Haggag et al. "Abstractive Arabic Text Summarization Based on Neural Networks", 2020.



- [10] Elaraby et al. "Arabic News Summarization Using Transformer-based Models", 2020.
- [11] Abooraig, and Al-Kabi, "Arabic Text Summarization Using Lexical Chains and Sentence Similarity", 2019.
- [12] Al-Sammaraie, and Al-Bakour, "Extractive Text Summarization for Arabic News", 2018.
- [13] Al-Smadi, and Al-Nawaiseh, "A Hybrid Approach for Arabic Text Summarization Using Clustering and Sentence Scoring", 2018.
- [14] Sunil Kumar, M. ., Sundararajan, V. ., Balaji, N. A. ., Sambhaji Patil, S. ., Sharma, S. ., & Joy Winnie Wise, D. C. . (2023). Prediction of Heart Attack from Medical Records Using Big Data Mining. *International Journal of Intelligent Systems and Applications in Engineering*, 11(4s), 90–99. Retrieved from <https://ijisae.org/index.php/IJISAE/article/view/2575>.
- [15] Al-Salman et al., "Arabic Text Summarization Based on Sentence Compression and Semantic Similarity", 2018.
- [16] El-Haj et al., "Abstractive Text Summarization for Arabic News", 2018.
- [17] Mahmoud, and Ezzeldin, "Arabic Text Summarization Using Sentence Ranking Based on Connectedness and Positional Weight", 2017.
- [18] Alhadlaq et al., "Extractive Summarization of Arabic Texts Based on Sentence Clustering", 2017.
- [19] El-Halees, and Al-Salman, "Arabic Text Summarization Using Lexical Chains", 2016.
- [20] Al-Sarem et al., "Arabic Text Summarization: A Survey", 2016.
- [21] Al-Twairish, and Al-Nafjan, "Arabic Text Summarization: A Survey", 2014.
- [22] AL-Khassawneh, Y. A., & Hanandeh, E. S. (2023). Extractive Arabic Text Summarization-Graph-Based Approach. *Electronics*, 12(2), 437.
- [23] Etaïwi, W., & Awajan, A. (2022). SemG-TS: Abstractive Arabic Text Summarization Using Semantic Graph Embedding. *Mathematics*, 10(18), 3225.
- [24] Wazery, Y. M., Saleh, M. E., Alharbi, A., & Ali, A. A. (2022). Abstractive Arabic text summarization based on deep learning. *Computational Intelligence and Neuroscience*, 2022.
- [25] A. Al-Sabbagh, M. K. Al-Subari, & A. Hamdan. (2017). Towards Building an Arabic Single-Document Summarization Dataset. In *Proceedings of the 8th International Conference on Computer Sciences and Convergence Information Technology (ICCIT)*.
- [26] Bouazizi, M., & Mezghani, A. (2015). Arabic Single Document Summarization Using Lexical Chains. *International Journal on Document Analysis and Recognition (IJ DAR)*, 18(3), 297-316.
- [27] <https://github.com/mkhzoumi/SAN>
- [28] Dhabliya, D. (2021). An Integrated Optimization Model for Plant Diseases Prediction with Machine Learning Model . *Machine Learning Applications in Engineering Education and Management*, 1(2), 21–26. Retrieved from <http://yashikajournals.com/index.php/mlaeem/article/view/15>.
- [29] Salleh, M. R., Shaari, A. H., & Zainuddin, R. (2018). "AraVec: A Set of Arabic Word Embedding Models Trained on a Large Corpus." *Journal of Physics: Conference Series*, 1049(1), 012039.
- [30] Abdelaal, K., Elaraby, M., Mourad, A., & Rafea, A. (2018). "Abstractive Arabic Text Summarization Using Deep Learning Techniques." In *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)* (pp. 1-4).
- [31] Al-Zaidy, R., & Mehdad, Y. (2019). MultiBooked: A Multidimensional Dataset for Arabic Text Summarization. In *Proceedings of the 4th Arabic Natural Language Processing Workshop* (pp. 16-24). Association for Computational Linguistics.
- [32] Stojanovic, N. . (2020). Deep Learning Technique-Based 3d Lung Image-Based Tumor Detection Using segmentation and Classification. *Research Journal of Computer Systems and Engineering*, 1(2), 13:19. Retrieved from <https://technicaljournals.org/RJCSE/index.php/journal/article/view/6>
- [33] Baly, R., Haddoud, M. H., Alsaied, T., & Alghamdi, R. (2018). MultiBooked: A multilingual multi-document summarization benchmark. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1362-1372).
- [34] <https://github.com/HLTCHKUST/MultiBooked>