Ingénierie des Systèmes d'Information
Vol. 31, No. 6, June, 2026, pp. 1949-1959
Journal homepage: http://iieta.org/journals/isi

# A Deep Reinforcement Learning Framework with Maximal Marginal Relevance-Based Reward Optimization for Query-Focused Extractive Multi-Document Summarization

Abeer AL Hassainy*, Wafaa AL Hameed

Software Department, College of Information Technology, University of Babylon, Babylon 51002, Iraq

Corresponding Author Email: abirhusseinj.sw@student.uobabylon.edu.iq



https://doi.org/10.18280/isi.310617

**Received:** 22 December 2025
**Revised:** 20 February 2026
**Accepted:** 3 March 2026
**Available online:** 30 June 2026

***Keywords:***
*query-focused summarization, extractive text summarization, deep reinforcement learning, Deep Q-Network, Maximal Marginal Relevance, BERT-based semantic representation, natural language processing*

**ABSTRACT**

Query-focused extractive summarization aims to identify the most informative sentences from multiple documents according to a specific user query while maintaining relevance, diversity, and coherence. Existing approaches often rely on static sentence ranking strategies, which may lead to redundant information and insufficient modeling of sequential sentence selection decisions. This study proposes a deep reinforcement learning framework with a Maximal Marginal Relevance (MMR)-based reward optimization strategy for extractive query-focused multi-document summarization. The framework formulates sentence selection as a sequential decision-making problem using a Deep Q-Network (DQN) agent. A multi-component reward function is designed by integrating query relevance, redundancy reduction through MMR, semantic coherence based on Bidirectional Encoder Representations from Transformers (BERT) similarity, and document coverage. The model was evaluated on the QuerySum benchmark dataset containing multi-document query-focused summarization (QFS) samples. Experimental results demonstrate that the DRL-MMR framework achieves ROUGE-1, ROUGE-2, and ROUGE-L F1-scores of 47.62, 22.46, and 31.75, respectively, together with a BERTScore of 83.68%. Ablation experiments further confirm that removing MMR or coherence components decreases summarization quality, indicating their complementary contribution to sentence selection. The findings show that integrating reinforcement learning with diversity-aware reward optimization can improve extractive summarization performance by balancing query relevance, information coverage, and redundancy control. The proposed framework provides a potential solution for intelligent information retrieval and automated text summarization systems.

## 1. INTRODUCTION

The rapid development of digital information has significantly increased the need for effective approaches that help users quickly access relevant information. As a solution to this problem, text summarization has emerged as a method for generating abridged representations of lengthy documents. Extractive query-based multi-document summarization can be described as the process of generating short summaries by selecting informative and relevant phrases or sentences from the original documents that directly respond to users' queries [1, 2]. This approach ensures that the summaries are linguistically accurate because the original content is preserved (Figure 1).

Unlike generic summarization, which aims to capture the overall content of a document, query-focused summarization (QFS) prioritizes the main sentences in a text that are directly relevant to the user's query. This reduces the time users need to spend reading full documents and helps them quickly identify relevant information, making QFS particularly useful for applications such as decision-support systems, question answering, and information retrieval [3].

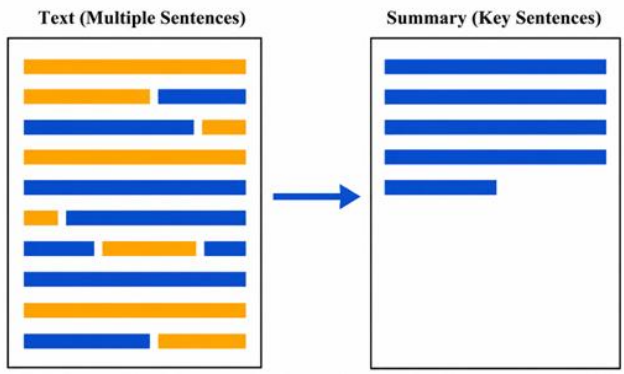


**Figure 1.** Summarization process [4]

Classical extractive summarization methods are largely based on optimization approaches, such as integer linear programming (ILP), graph-based methods, and techniques that use human-designed statistical features, such as sentence length or position, to identify significant sentences within documents [5]. These techniques can capture surface-level significance, but they often struggle to model the semantic

relationships between user queries and sentences effectively. As a result, the generated summaries may contain redundant sentences or information that is only loosely related to the query, thereby reducing their readability and overall relevance [6].

Recent advances in deep learning have improved text representation through neural embedding models, including transformer-based architectures such as Bidirectional Encoder Representations from Transformers (BERT), thereby enabling a better semantic understanding of queries and documents. These models have been successfully applied to various natural language processing (NLP) tasks, including document summarization [7]. However, many deep learning-based summarization methods still rely on static ranking techniques that assess sentences independently. These methods do not adequately capture the sequential nature of sentence selection in summarization, in which each selected sentence affects the diversity and relevance of the remaining candidates.

Deep Reinforcement Learning (DRL) has emerged as a promising solution to these challenges and has transformed several NLP applications, including extractive summarization [8, 9]. In Reinforcement Learning (RL)-based frameworks, the summarization task is modeled as a sequential decision-making problem. DRL reduces dependence on handcrafted features by using embeddings to represent documents, sentences, and words. An agent learns to select sentences by optimizing a reward function that assesses the relevance and quality of the summary [10].

Despite these advances, existing RL methods for QFS still face several significant challenges. In many cases, models prioritize relevance to the query but do not adequately control redundancy among the selected sentences. This results in summaries that may contain redundant or highly similar information, thereby reducing their coherence and informativeness [11, 12]. Furthermore, traditional diversification mechanisms such as Maximal Marginal Relevance (MMR) are well known for their ability to promote diversity and reduce redundancy among selected sentences. However, MMR alone cannot learn adaptive sentence-selection policies from data and therefore does not fully exploit the contextual relationships between documents and queries [13].

To address these limitations, this research proposes a hybrid framework that combines the MMR technique with Deep Q-Network (DQN)-based RL for extractive query-based summarization (EQBS). In the proposed system, DQNs demonstrate strong potential for learning effective sentence-selection policies. By interacting with the environment, the agent can gradually learn which sentences should be included in the summary to maximize query relevance and overall summary quality. Meanwhile, the MMR technique is integrated to enhance diversity and explicitly reduce redundancy among the selected sentences.

Unlike previous approaches that rely on single-document datasets, the proposed system was evaluated through extensive experiments on the query-oriented multi-document QuerySum dataset using ROUGE metrics. The evaluation demonstrates that the proposed system provides stronger query relevance and diversity while maintaining fluency and coherence. The reward function enables the agent to generate logically consistent and semantically rich summaries.

The major contributions of this study can be brief as follows:

•A DRL framework depend on DQN is suggested for extractive query-based multi-document summarization, where the agent learns an optimal sentence selection policy guided by the query.

•The MMR technique is combined into the reward function to minimize redundancy and enhance the diversity of the chosen sentences through the summarization process.

•A BERT-based semantic coherence component is integrated to guarantee logical consistency between sequentially chosen sentences in the created summary.

The remainder of this paper is organized as follows: Section 2 reviews recent state-of-the-art works in extractive, abstractive, and RL-based summarization. Section 3 describes the research questions for this study. Section 4 describes the proposed system. Section 5 details the datasets and experimental results, and evaluation. Section 6 presents quantitative and qualitative results, including ablation studies and human evaluations. Finally, Section 7 concludes the paper and outlines future directions.

## 2. RELATED WORK

Studies on automatic text summarization have developed significantly over the past decades. Current research can generally be classified into several directions, such as redundancy reduction techniques (e.g., MMR), QFS methods, and RL-based summarization.

### 2.1 Redundancy reduction technique

The paper [14] natural language processing as a knowledge-intensive (KI) model that eliminates the basis on pre-existing document groups. This system comprises two components: a summarization controller and a retrieval module. The retrieval module is Dense Passage Retrieval (DPR), which automatically searches using methods like BM25 for potentially relevant documents from a large-scale knowledge corpus. Then the summarization controller combines a multi-step process approach with a large language model (GPT-3.5). A few-shot prompting strategy to first extract query-relevant information and then generate a concise, non-redundant summary. The limitation of this study is that the model's performance decreases slightly when operating in a large open-domain corpus.

This research produces an interactive query-based summarizing system to enhance the relevance and quality of summaries for scientific documents [15]. The suggested system incorporates user interaction at several stages to enhance the summary, which integrates user participation by introducing a chance for the user to use the RAKE algorithm. Using the GA algorithm, the system can expand the chosen sentences; the final summary is created by aggregating the set of user-preferred and expanded sentences and ranking them using the MMR technique. It evaluated the effect of interaction with the user through various methods. 45 AI graduate students tested the proposed model by applying it to 72,000 scientific documents from the Web of Science (AI category).

### 2.2 Query-Focused Summarization

Yao et al. [16] introduced an RL system for QFS, which replaces conventional supervised learning with a policy-gradient training model, allowing the paradigm to enhance sequence-level rewards rather than token-level cross-entropy

objectives. This method addresses the well-known incompatibility between RL and the teacher force, enabling stable gradient propagation across token sampling. The suggested system combines multiple reward signals—including a novel semantic similarity reward derived from custom passage embeddings trained via the Cluster Hypothesis and BLEU and ROUGE—which significantly optimizes semantic query relevance. This study is limited by several factors, including the need for foundational computational resources, such as multi-GPU hardware and long training times, which prevented comprehensive hyperparameter tuning, according to action-space exploration and trust assignment difficulties.

The RL training through long text sequences is still challenging. Zhang et al. [17] introduced the Qontsum, a contrastive learning model to optimize the relevance between the user's query and the created summary. Different from conventional approaches, which are based on attention mechanisms only, Qontsum explicitly distinguishes between relevant and irrelevant document chunks. This framework faces several limitations; it cannot explicitly balance between relevance, redundancy, and difficulty in handling general or very specific queries. Also, this proposed model is based on contrastive fine-tuning but does not handle computational efficiency or adaptability in multi-document or large-scale QFS scenarios.

Lamsiyah et al. [18] proposed to leverage transfer learning from pre-trained sentence embedding models (uSIF, USE-DAN, and USE-Transformer) to represent documents' sentences and users' queries. Furthermore, BM25 and the semantic similarity function are linearly combined to retrieve a subset of sentences based on their relevance to the query. Finally, the MMR criterion is applied to re-rank the selected sentences by maintaining query relevance and minimizing redundancy. Experiments are conducted using three standard datasets from the DUC evaluation campaign (DUC 2005–2007). This proposed approach is limited by its reliance on fixed-length summaries (e.g., 250 words). The evaluation is limited to standard benchmark datasets (DUC 2005–2007), lacking sentence-level coherence modelling and the absence of fine-tuning the pretrained models. The linear combination of BM25 and semantic similarity scores, although effective, is manually weighted and lacks a learned optimization process that could better capture complex interactions between lexical and semantic relevance. In future work, I plan to investigate the potential of the newest models, T5 and GPT-3-3-3, for the text summarization task. And explore transfer learning abilities from pre-trained models for summary generation.

## 3. RESEARCH QUESTIONS

The purpose of this study is to respond to the research questions, which are:

1. Is the selected sentence for generating a summary semantically coherent with the already selected ones and relevant to the query?
2. What is the contribution of the proposed system in summarizing multiple documents logically and consistently to answer user queries and identify whether a retrieved document is relevant and improve the search for and access to critical and salient information efficiently?
3. How can we design a DRL framework to effectively estimate the $Q$-value function with coverage, non-redundancy, and semantic relatedness among selected units (sentences) for EQBS? Because in DRL, the $Q$-value function estimates the expected long-term reward for taking an action (e.g., selecting a sentence) in a given state.

## 4. PROPOSED SYSTEM

### 4.1 Problem formulation

In this research, for the EQBS process, multiple documents $\{D_1, D_2, ..., D_n\}$ and query $Q$ are given as input; the combination of $D$ and $Q$ represents an environment where each document includes m sentences $\{S_1, S_2, ..., S_m\}$, which is the possibility of the source of an action. The selected sentences and their relationships with the query represent the current state ($S$). An action ($a$) will be either selecting or skipping a sentence based on the current state. The reward function ($R$) considers relevance to the query based on the cosine similarity measure, novelty to avoid redundancy, and coherence. The main goal of building the system is to create an extractive query-based summary $E$ by selecting $m$ sentences from $D$, where $E = \{se_1, se_2, \ldots, se_m\}$ and ($m < n$). This summary must be prominent, should not have repeated sentences, should be relevant to the query, and should be readable.

### 4.2 Deep reinforcement learning for extractive query-based summarization

Inspired by extractive summarization approaches based on traditional and deep learning, the research proposes a DRL agent for an EQBS process. The proposed system architecture is illustrated in Figure 2. This system incorporates several NLP techniques; each is crucial in ensuring that the generated summary is salient and most relevant to the query, novelty, coverage, and informativeness summaries, maintaining coherence. Building this system includes many carefully designed stages. Initially, the input and preprocessing stage: In this step, the input for the training and testing steps of the system will be a corpus of data that is a set of documents $\{D_1, D_2 \ldots D_n\}$ with multiple queries $Q$ and multiple reference summaries (human-written also known as "gold summaries"). At this stage, the proposed system applies the preprocessing step, which seeks to transform the raw document text into a format the machine can understand and process, making it more interpretable for subsequent stages. This step involves several techniques, such as noise removal to clean the text by filtering out stop words, removing HTML tags, and removing undesired characters and parentheses, email addresses, URLs, etc., to reduce the complexity of the input sentences and queries (raw data) by retaining only semantically relevant content and converting it into a format understandable by a machine, followed by the reduction of extra whitespace and normalization. It is important to convert the text to lowercase to ensure uniformity for the best semantic meaning, which is necessary in the EQBS task. After the preprocessing step, in this stage, the system uses an NLP library (spaCy model) for tokenizing the cleaned text (the output from the preprocessing step) into sentences. Sentence tokenization is one of the main issues. The segmentation process is also referred to as boundary detection. Unlike classical methods like regular expressions, which are based on NLTK's Punkt tokenizer, or splitting, spaCy leverages capitalization and dependency parsing in integration with punctuation to determine sentence

boundaries more reliably. This approach ensures that each document is decomposed into cohesive sentence units while maintaining semantic context. Therefore, it provides high-quality input for clustering and RL-based summarization tasks and subsequent embedding. Then, each sentence and query are embedded. In this step, the proposed system converts each token (clean sentences, clean queries) into numerical vector representations using a pretrained BERT model, which was implemented to obtain contextual representations of text. Since BERT is bidirectional, this feature makes it possible for BERT to efficiently learn the "context" of each word by examining its neighbouring words, where the model utilizes an attention method for focusing on significant words, guaranteeing that semantic similarities can be effectively caught even when there is lexical variation between the document and the query. This results in a powerful base for similarity-based sentence choice and summarization based on a query. The summarization environment with reward function design is defined by employing a DRL where the summarization process (i.e., the sentence selection process) is modeled as a Markov Decision Process (MDP) with concepts (state representation, action space, and reward function).

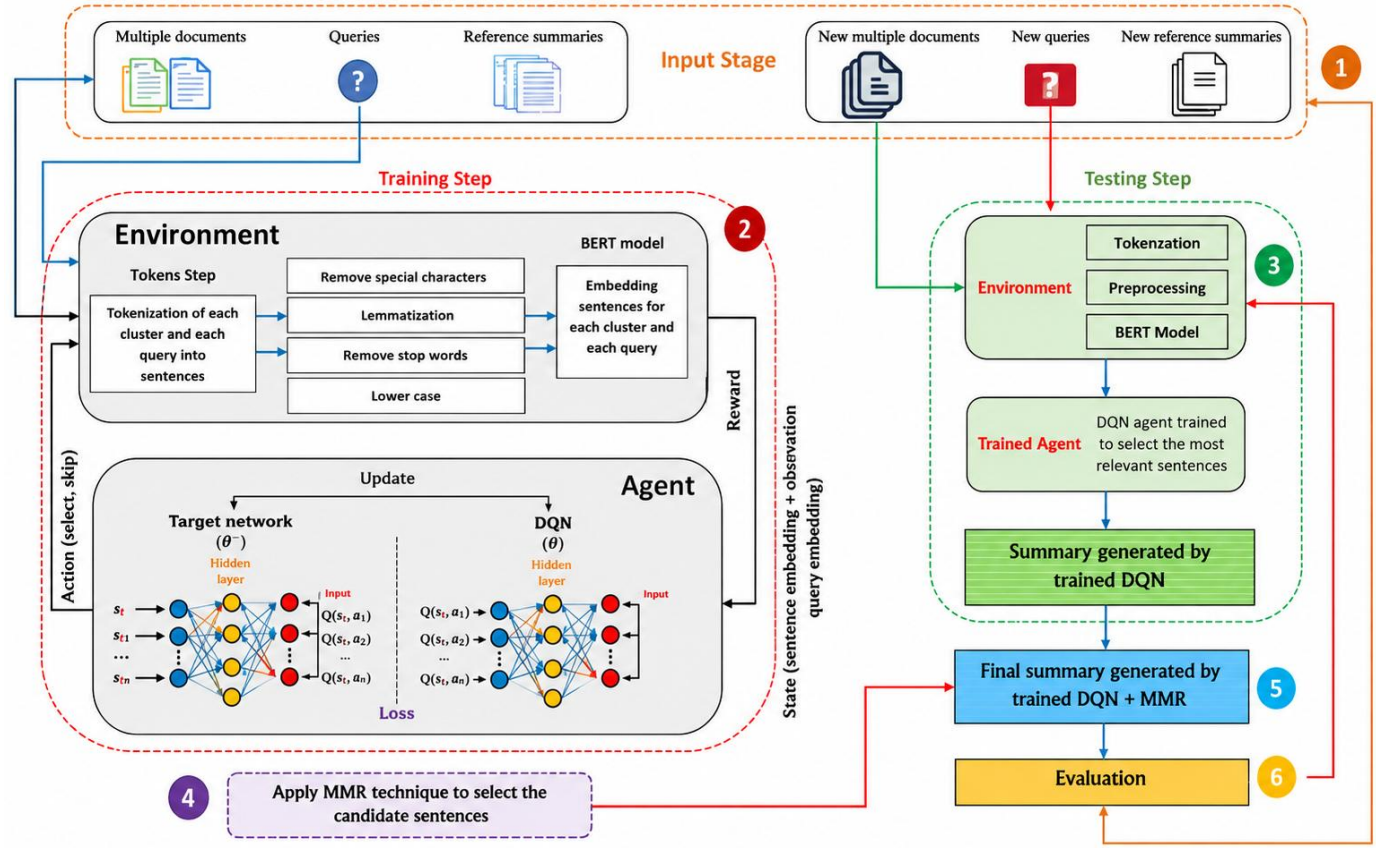


**Figure 2.** Proposed system (DRL+MMR)
Note: DRL = Deep Reinforcement Learning, MMR = Maximal Marginal Relevance.

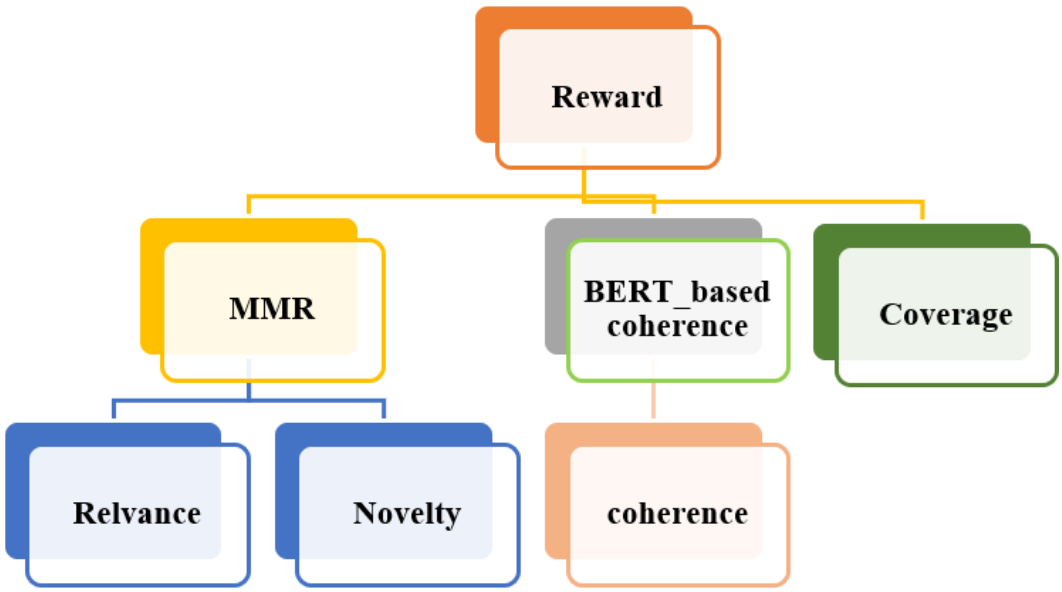


**Figure 3.** Reward components

The combination process between the selected sentence vector and query vector, resulting from the embedding stage, contributes to constructing an environment that simulates how an agent interacts to generate a summary. The constructing agent observes the current state represented by the integration of the current sentence (selected sentences or remaining candidates) embedding and the query embedding, allowing an agent to make context-aware selection decisions from the action space. The environment defines two potential actions: select or skip a sentence and transition through the document sequentially, updating the agent's state after every decision taken. During this process, the agent accumulates rewards until all sentences are processed or a maximum summary length is reached. This proposed framework enables the DQN agent to learn an optimal extraction policy that maximizes semantic quality and cohesion while minimizing redundancy, thereby distilling high-quality summaries based on queries. Reward function design is one of the most important contributions of the proposed system. Because it helps an agent learn how to select sentences that are coherent, most relevant, and non-redundant, corresponding to the query. The reward function is carefully constructed as a multi-component that balances three basic measures of summary quality that are relevance, coverage, and coherence, while penalizing redundancy, as shown in Figure 3. This model implements the MMR technique with a BERT-based coherence model and a coverage measure in building the reward function, as in Eq. (1), that will be given to an agent during training (at every time the DQN agent chooses a sentence, the reward will be calculated). Each part is weighted based on its contribution to the total summary quality.

$$r_t = \underbrace{(\lambda_1 \times Relevance + \lambda_2 \times Novelty)}_{MMR\ technique} + \underbrace{(\lambda_3 \times Coherence)}_{\text{BERT-based coherence}} + \lambda_4 \times Coverage \quad (1)$$

where, $r_t$ denotes the reward value at time step $t$. $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are weighting parameters that control the contribution of each component in the reward function.

In this proposed system, the priority given to MMR (0.60) is to maintain relevance and diversity, while (0.25) is provided for BERT-based coherence to preserve the semantic fluency. Finally (0.15) assigned for coverage scores, ensuring the information integrity. The reward weights were assigned according to the proper significance for each component in QFS, prioritizing relevance and redundancy reduction, followed by coherence and coverage.

The $\lambda_1 \times Relevance + \lambda_2 \times Novelty$ represents the implementation of MMR, which aims to handle some challenges faced by generative extractive summarization tasks, such as relevance to the query (measured using cosine similarity between the query embedding and sentence embedding), as described in Eq. (2).

$$Relevance(S, Q) = cos(embedding(S), embedding(Q)) \quad (2)$$

where, $S$ denotes the candidate sentence. $Q$ represents the query.

Diversity (novelty) is achieved by reducing duplicates and ensuring that the selected sentences are distinct from those previously selected for the generated summary as illustrated in Eq. (3).

$$Novelty(s, S_{selected}) = - \max_{s' \in Sselected} cos(embedding(s), embedding(s')) \quad (3)$$

where, $s$ represents the candidate sentence. $S_{selected}$ denotes the set of sentences that have already been selected in the summary. $s'$ represents a sentence belonging to the selected summary set $S_{selected}$. $embedding(\cdot)$ denotes the vector representation generated using the BERT model. $cos(\cdot)$ represents the cosine similarity function.

The MMR guarantees that the final summary is comprehensive and varied. It captures various aspects of the query without repeating information, as described in Eq. (4). This process ends when the required length of the generated summary is reached.

$$MMR(s) = \underbrace{\lambda \times Similarity(s, Query)}_{Relevance(S,Q)} - \underbrace{(1-\lambda) \times \max_{s' \in S} Similarity(s, s')}_{Novelty(news,Sselected)} \quad (4)$$

where, $s$ represents the candidate sentence. $Q$ denotes the input query. $S_{selected}$ shows the set of sentences already selected in the summary. $s'$ is a sentence in the selected set $S_{selected}$. $Similarity(\cdot)$ denotes the cosine similarity between sentence embeddings generated using the BERT model. $\lambda$ represents the balancing parameter controlling the trade-off between query relevance and redundancy reduction.

In QFS, guaranteeing that chosen sentences are both non-redundant and relevant is an essential objective; for this reason, it has the highest weight. During design, the reward function also takes into consideration the coherence component (0.25), introduced to encourage semantic consistency between consecutively chosen sentences, enhancing the logical flow and readability of the created summaries. The BERT-based coherence model in Eq. (5).

$$coherance(s_{t-1}, s_t) = \frac{1}{n-1} \sum_{i=1}^{n-1} cos\ (s_{t-1}, s_t) \quad (5)$$

where, $s_t$ represents the current candidate sentence selected at time step $t$. $s_{t-1}$ denotes the previously selected sentence in the summary. $embedding(\cdot)$ represents the vector representation generated using the BERT model. $\cos(\cdot)$ denotes cosine similarity used to measure semantic coherence between sentences. *n* represents number of sentences already selected in the summary.

Finally, the coverage component (0.15) guarantees that significant query-related information appears in the summary without too constraining the sentence chosen issues. Were calculated by comparing the summary embedding to the document embedding for calculating coverage as described in Eq. (6).

$$Coverage(S, D) = cos\left(mean_{embedding(S)} mean_{embedding(D)}\right) \quad (6)$$

$$mean_{embedding(S)} = \frac{1}{|S|} \sum_{si \in S} embedding(si)$$

$$mean_{embedding(D)} = \frac{1}{|D|} \sum_{di \in D} embedding(\text{dj})$$

where, $S$ represents the set of sentences selected in the generated summary. $D$ denotes the set of sentences in the source documents. $embedding(\cdot)$ represents the BERT-based sentence embedding representation. $mean_{embedding(S)}$ represents the average embedding vector of all sentences in the generated summary. $mean_{embedding(D)}$ represents the average embedding vector of the source document sentences. $cos\ (\cdot)$ denotes the cosine similarity function used to measure semantic similarity

These weights are considered a priority, depending on design selection, where relevance and redundancy control are emphasized first, followed by coherence and coverage.

DQN Agent Design: In the DRL model, an agent is built using two deep neural networks, a DQN and a target network with experience replay. The DQN agent is used to learn an optimal policy for selecting sentences, thereby predicting the long-term cumulative reward for each possible action, as illustrated in Eq. (7)

$$cumulative\ reward = (Rt)r_t + \gamma r_{t+1} + \gamma 2 r_{t+2} + \cdots \quad (7)$$

where, $R_t$ denotes cumulative reward starting from time step $t$. $r_t$ denotes immediate reward received at time step $t$. $\gamma$ denotes discount factor that controls the importance of future rewards ( $0 \leq \gamma \leq 1$ ).

This cumulative reward reflects the overall quality of the created summary in terms of redundancy reduction, coverage, relevance, and coherence.

The DQN represents a basic decision-making technique in the suggested model, which employs a deep neural network to approximate the $Q$-value function to generate informative features to represent the states of the RL Bellman model. DQN is used when the action space or states are large (2017). The agent learns which sentences maximize the reward, balancing query relevance and information diversity. The agent utilizes experience replay and policies to balance exploration of new sentence combinations and exploitation of known good policies. At the training step, an agent periodically selects sentences, updates the state, and receives a reward. A target network is periodically synchronized with the DQN network through many episodes and is used to stabilize learning by reducing oscillations in $Q$-value updates. At each step, experience replay is employed to store past transitions as tuples (state(s), action(a), reward(r), next state(s'), done), which provide samples for training an agent. During the training stage, an agent periodically selects sentences, updates the state, and receives rewards. This allows the model to work in a sequential decision-making process and to generate high-quality extractive summaries. Training the DRL model and exploration: The training approach of the DQN agent proceeds by interacting with the specific environment repeatedly; an EQBS is handled as a sequential decision-making process. At every step, the agent observes the current state vector (a concatenation of the sentence embedding and the query embedding). The DQN architecture consists of fully connected layers with non-linear activations that approximate the optimal action-value function $Q(s,\ a)$, This results in $Q$-values representing two possible actions (selecting this sentence or skipping it). It captures a reward based on a multi-component function that integrates BERT-based coherence, MMR for controlling both the relevance and redundancy, and query coverage. These rewards help the agent select sentences that are aligned with the query, contextual, and informative. To guarantee stable learning, a separate target network is preserved and synchronized periodically with the main network. The $Q$-network is updated iteratively using experience tuples $(s, a, r, s')$ to enhance convergence stability. These tuples are stored in the replay buffer after each experience and sampled during the training process. An agent, when interacting with an environment during the learning process, balances exploitation and exploration using an ε-greedy strategy, where each chosen sentence is evaluated by the reward function, enabling the agent to progressively move from random exploration to policy-driven action selection based on learned $Q$-values. Testing stage: This stage of the proposed system involves the implementation of the trained (DQN) agent, forming a final concise extractive summary for each query by concatenating selected relevant sentences tailored to the query. The DQN agent transitions from a learning environment to a completely deterministic decision-making model. Where, after training the agent (in the training stage), input a new corpus of a dataset with the same structure as that used in training the agent (multiple documents, multiple queries, and multiple reference summaries) for an initialization step to apply preprocessing, tokenization, and embedding stages. Unlike training, at this stage, an agent exploits the state space instead of exploring it. The agent is applied only in deduction mode. It selects the action with the highest $Q$-values acquired during the training stage, guaranteeing a purely exploitative behavior, leveraging the learned policy to assess every sentence corresponding to the query. Moreover, no exploration, no weight updates, and no backpropagation are performed. An environment sequentially provides sentence embeddings concatenated with the query embedding, which was achieved using the BERT model for the trained DQN agent to interact with an environment. At every step, the agent provides $Q$-values for two possible actions and determines whether to choose or skip a sentence depending on the predicted coherence and relevance. This decision procedure continues until either all sentences are processed or the maximum summary length is reached. The essential components for stabilizing learning during the training stage are the target network and experience replay, which remain inactive during testing. Therefore, the agent is treated as a constant decision model, continuously implementing its understanding of coverage, relevance, and coherence to generate the best extractive query-based summary. Essentially the trained agent assignment is represented as an intelligent selector, autonomously filtering the most informative sentences without any environmental feedback or further adaptability. Summarization-generation stage: This stage involves a trained DRL agent working as a deterministic decision-maker to form a final concise extractive summary by concatenating selected relevant sentences tailored to the query. It assesses every sentence by ensuring that the selected sentences are different from the sentences that were selected before to be included in the generated summary. It captures various aspects of the query without repeating information by utilizing the learned $Q$-function and greedily choosing just the most relevant sentences. This process will end when the required length of the generated summary is reached. Compared to traditional supervised summarization models, this DQN-based system does not require aligned, labeled data for every possible summary. Instead, it learns from rewards based on its own decisions, making it more flexible and adaptive to unseen queries and document sets. The agent applies a select-or-skip policy sequentially to create an extractive summary.

## 5. EXPERIMENTAL RESULTS AND EVALUATION

This section describes the evaluation of the proposed system, explaining the results and their corresponding outcomes. Refers to the benchmark datasets that depend on it in the evaluation process.

### 5.1 Dataset and queries

The suggested system utilizes the QuerySum dataset. It is a benchmark for multi-document QFS issues. QuerySum, including 27,041 high-quality samples, each of which contains a natural-language query (Q), a human-validated summary (reference summary), and up to 10 related multiple-source documents, with an average of 5.5 per sample, retrieved primarily from Wikipedia. QFS simulates real-world requirements, where the user introduces a question, and the proposed system retrieves the condensed and relevant information. Unlike classical summarization datasets, QuerySum seeks to generate summaries dependent on user queries, which are complex questions across multiple documents. Focuses on non-factoid queries, such as "why," "what," and "how" queries that demand a deeper understanding and synthesis of information. This makes it more challenging and realistic compared to factoid-based QA (query answering) datasets. The dataset [19] is built using seed queries that were extracted from Answers.com and extended via Google Search to form query clusters, which are further annotated manually to define semantic relationships (synonymous, related, or unrelated). Used Wikipedia as the main source for documents. This structure supports deeper modeling of query intent and semantic generalization. Furthermore, a query is a fundamental part of the input stage. It provides the intent or context that helps in choosing sentences from documents to create extractive-query-based summarization. This query will be either extracted automatically using different approaches for generating it, for instance, Text Rank for keyword extraction or generating short queries, or written by a human (i.e., written by users or experts in the domain). In the proposed system, a specific query written by the user is used, aiming to generate the most relevant and diverse document, but it may not exactly match the reference but is highly accurate and relevant to the query. Furthermore, human evaluation can be used to assess the suggested system by taking into consideration the relevance to the query, the readability, and the diversity.

### 5.2 Experimental setup

To evaluate the execution of the suggested DRL + MMR model for EQBS, employ specific users' queries and a benchmark multi-document summarization dataset that includes documents and human-written highlight summaries. Every document was split into tokens using the NLTK sentence tokenizer. A preprocessing stage, containing punctuation removal and lowercasing, was applied. Utilize the 'all-MiniLM-L6-v2' variant of Sentence-BERT for query representation and sentence creation, generating semantically rich embeddings that simplify MMR-based diversity and relevance scores. Evaluation is implemented to learn policies for selecting sentences. The environment was formulated as an MDP, including a combination of the query embedding and the embedding of selected sentences referring to the state. Then, an action represented selecting sentences from the unselected aggregation or skipping it. Finally, the reward is designed to be a consideration for many factors, such as the coherence score and the redundancy penalty, by using MMR and relevance to the query.

## 6. RESULTS AND DISCUSSION

To validate the effectiveness of the suggested system, we a comparison with previous focused summarization approaches on the same QuerySum dataset utilized in our experiments to guarantee a fair comparison under the same experimental conditions. The results of implementing the proposed system are thus shown in Table 1, which shows the efficiency performance of the system compared with previous extractive baselines on the QuerySum dataset in creating redundancy, relevance, and a readable summary based on evaluation metrics ROUGE-1, ROUGE-2, ROUGE-L, and BERT score value. This improvement is attributed to the integration of DRL, MMR, and BERT-based semantic alignment and an RL reward that combines coherence, novelty, and relevance.

The results in the same table are reported from the original publications and are presented only for general reference when compared to the other dataset (CNN Daily Mail). Since these models were assessed on various datasets, their ROUGE scores are not directly comparable with those obtained on the QuerySum dataset. The proposed system uses a training DQN agent with a reward function built to take into consideration convergence, coherence, relevance, and novelty to generate an EQBS agent trained to interact with an environment represented by multiple documents and a specific query, which are initially tokenized into sentences using the spaCy model, as shown in Figure 4. These sentences are pre-processed by removing special characters, lemmatizing, and converting to lower case, and finally utilized by the BERT model to convert these sentences and the query into numerical vectors.

**Table 1.** Results of implementing the proposed system compared with the previous

| Approaches | Datasets | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|---|
| | | F-Measure | F-Measure | F-Measure |
| Proposed system (DRL-MMR) | QuerySum | 47.62 | 22.46 | 31.75 |
| Qontsum | QMSum | 45.76 | 14.27 | 24.14 |
| IRL | CNN/DailyMail | 44.61 | 20.14 | 39.04 |
| | CNN/DailyMail | 39.3 | 15.8 | 35.2 |
| DQN-CNN-RNN | DUC 2002 | 39.4 | 16.1 | 35.6 |
| DQN-RNN-RNN | CNN/DailyMail | 45.9 | 22.3 | 42.5 |
| | DUC 2002 | 46.4 | 22.7 | 42.9 |
| RNES w/o coherence | CNN/Daily Mail | 41.2 | 18.8 | 37.7 |
| RNES w/ coherence | CNN/Daily Mail | 40.9 | 18.6 | 37.41 |
| Query-Based Unsupervised Extractive Method (QBUEM) | Debatepedia | 34.67 | 6.54 | 28.05 |

Note: DRL = Deep Reinforcement Learning, MMR = Maximal Marginal Relevance.

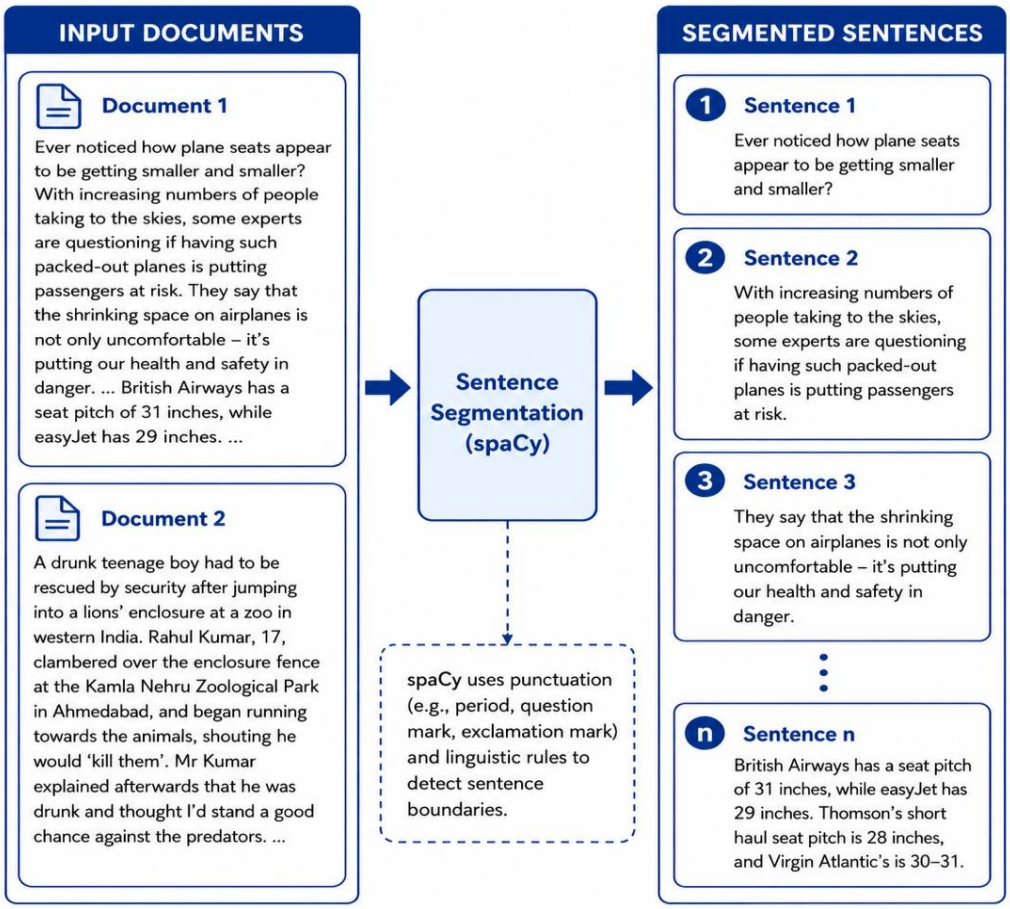


**Figure 4.** Tokenization process result

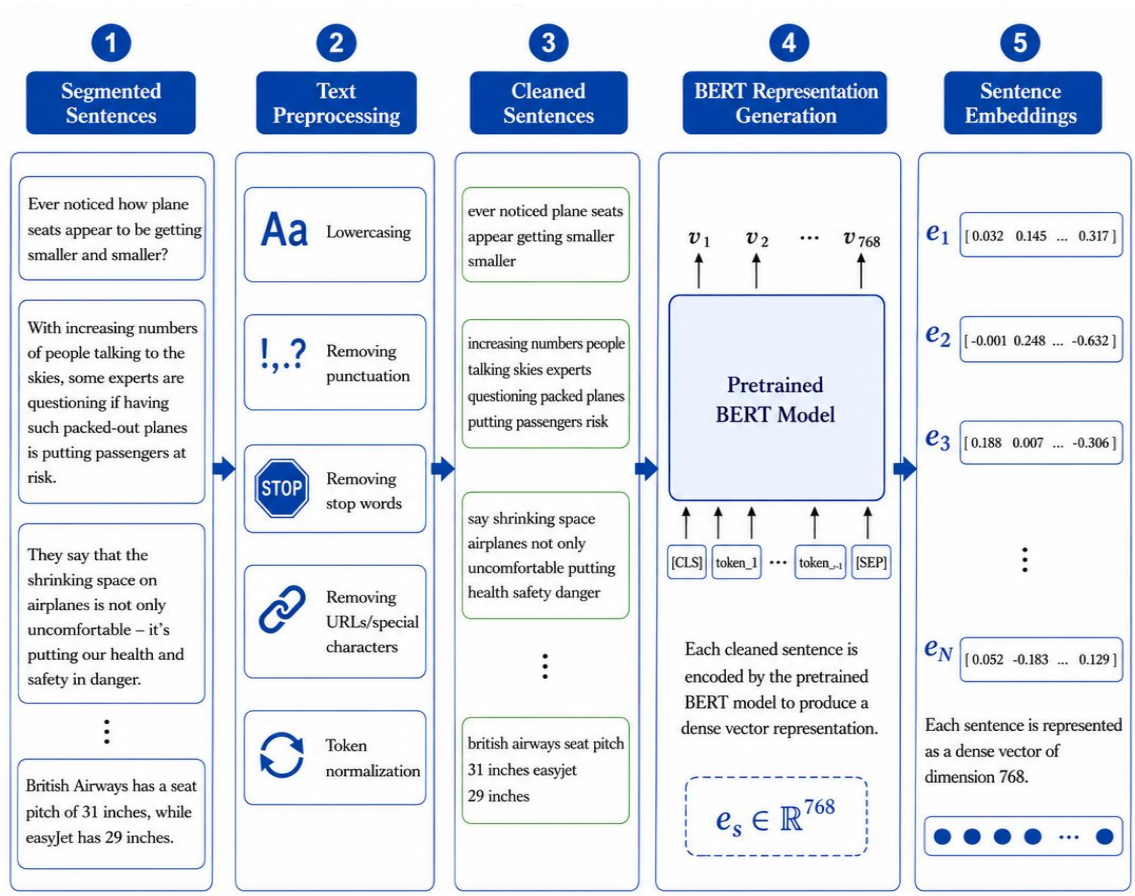


**Figure 5.** Sample of episode during training

**Table 2.** Results of implementing the proposed system

| Metric | Precision | Recall | F-Measure |
|---|---|---|---|
| ROUGE-1 | 61.6496 | 38.79 | 47.62 |
| ROUGE-2 | 29.17 | 18.26 | 22.46 |
| ROUGE-L | 41.10 | 25.86 | 31.75 |

The combination of the sentence embedding vector (768 in length, because using the BERT model) with the query embedding vector (also 768) results in a state (1536 from 768 sentence embeddings + 768 query embeddings) that an agent observed from an environment. Since no sentences have been chosen yet, an environment is initialized with an empty summary. This step reduced semantic noise by enabling subsequent models to assess sentence relevance, redundancy, coherence, and contextual fit in relation to user queries and ensured that summarization decisions remained aligned with thematic intent. Overall, the tokenization process successfully demonstrates its essential role in the architecture of the proposed framework. At each episode of training, the agent is trained to select the sentence or skip it based on the reward function that was built to help the agent to decide, as explained in Figure 5. This figure represents a sample about training, including the query that is input for training, and explains how the DQN network predicts the *Q*-value to represent the selected action (0 for skip the sentence and 1 for select the sentence). In the figure, "select sentence" represents the current state that is observed by an agent, and "next sentence" refers to the next state that an agent observes and will be input for the target network to predict the *Q*-value, which helps to balance the training episode After training, to improve diversity, reranked the selected sentences using MMR as a post-process. This step helps to answer the second research question. Focusing the reward function on the key factors of coverage and query relevance helped an agent select the most relevant sentences that are associated with the user's requirement, which helps to answer the first research question.

A comprehensive evaluation was conducted using both lexical and semantic metrics. Specifically, ROUGE-1, ROUGE-2, and ROUGE-L F1-scores were employed to measure n-gram overlap and structural similarity between the generated summaries and hu-man-written reference highlights. These metrics capture essential aspects such as unigram recall (content coverage), bigram coherence (phrase-level fluency), and longest common subsequence (overall structural alignment) as described in the Table 2 demonstrating the efficiency of the proposed system in generating the EQBS.

BERT score value (83.684%) enhancements illustrate that the created summaries caught heavy semantic meaning compared to simple lexical overlaps. Referring to the fact that the created summary has better concatenation alignment with the golden summaries.

As explained in Table 3, the chosen sentences were syntactically and contextually suitable and the suggested approach successfully chooses sentences that capture the main query-related information while minimizing redundancy and preserving reasonable coherence.

**Table 3.** Generated summary compared with reference summary

| Queries | Generated Summaries | Reference Summaries | ROUGE | | BERT Similarity Score | |
|---|---|---|---|---|---|---|
| | | | R1 | R2 | R-L | |
| Champions PSG Palermo | Palermo forward Paulo Dybala is reportedly being tracked by several major European clubs, including Manchester United, Juventus, Paris Saint-Germain, Inter Milan, Chelsea, and Arsenal. Palermo is expected to demand approximately £36 million for the player, which may lead to a bidding contest among the interested clubs. | Paulo Dybala has attracted interest from several leading European clubs, including Manchester United, Juventus, Paris Saint-Germain, Inter Milan, Chelsea, and Arsenal. Inter Milan manager Roberto Mancini was reportedly seen watching him play, while Palermo is believed to be seeking approximately £36 million for the forward. | 44.83 | 24.56 | 27.59 | 89.21 |
| Current salt guidelines | Current dietary guidelines recommend that a typical healthy adult consume no more than 2,300 mg of sodium per day. A stricter limit of 1,500 mg is commonly recommended for older adults and individuals with certain health risks. However, some researchers have questioned whether very low sodium intake may also be associated with adverse health outcomes. | Current federal guidelines recommend a daily sodium limit of 2,300 mg for healthy adults and approximately 1,500 mg for people over the age of 50 or those at greater health risk. Some studies suggest that consuming less than 3,000 mg per day may also be associated with health concerns, while average daily intake is approximately 3,500 mg. | 65.00% | 47.46% | 51.67% | 90.67% |
| Cricketers rescued Pakistani | Zimbabwe reportedly agreed to visit Pakistan for a cricket series despite continuing security concerns. International cricket in Pakistan had been limited since the 2009 attack on the Sri Lankan team bus in Lahore. During that attack, Sri Lankan players were rescued from Gaddafi Stadium and transported to safety. | Zimbabwe reportedly agreed to play a cricket series in Pakistan, which had not hosted major international cricket since the 2009 attack on the Sri Lankan team bus in Lahore. The incident raised serious security concerns after gunmen attacked the team convoy and killed several people. | 37.74% | 15.38% | 20.75% | 87.87% |

Note: BERT = Bidirectional Encoder Representations from Transformers.

## 7. ABLATION STUDY

To evaluate the contribution of each part in the proposed system, we conducted an ablation study by removing one component at a time and preserving the other components without change, resulting in three arrangements: (1) a DRL baseline without coherence or MMR; (2) integrated DRL with MMR only; and (3) combined coherence scoring only with DRL as shown in Table 4.

**Table 4.** Results of the ablation study for different model variants using ROUGE and BERTScore

| Model Variant Approaches | R1 | R2 | R-L | BERT Similarity Score |
|---|---|---|---|---|
| Proposed model (DRL+MMR) | 61.6496 | 61.6496 | 61.6496 | 83.68 |
| Proposed model removing MMR | 48.22 | 20.9 | 38.9 | 85.5 |
| Proposed model removing BERT coherence | 44.9 | 17.1 | 28.5 | 83.5 |

Note: DRL = Deep Reinforcement Learning, MMR = Maximal Marginal Relevance, BERT = Bidirectional Encoder Representations from Transformers.

Isolating the MMR module leads to an increase in redundancy. While removing the coherence component, it reduced sentence-level fluency and logical consistency. The outcomes of the full framework, incorporating all components, achieved the highest BERTScore and ROUGE values, confirming that each component plays a crucial and complementary role in the total performance of the system.

## 8. CONCLUSION

This research suggested a system for EQBS that integrates DRL with MMR to create diverse, concise, coherent, and relevant summaries from multiple documents in response to user queries. The sentence selection task is formulated as a sequential decision-making issue framed by the MDP. Based on learned policy, an agent selects sentences iteratively. Utilization of MMR in this system assisted in balancing redundancy and query relevance. It helps not only to cover the main topics of the query but also to preserve diversity between chosen sentences. The experimental results explain that the suggested system (DRL + MMR) is more efficient compared with the baseline approaches, such as random selection, Text Rank, and Lex Rank, according to BERT Score, ROUGE-1, ROUGE-2, and ROUGE-L metrics. It is necessary to refer to the proposed system, which carries out a higher semantic similarity to the query and a decreased redundancy rate. Demonstrating the performance of the combination of diversity control through inference and training. Moreover, the results showed that the closely designed reward function, taking into consideration novelty, coverage, relevance, and coherence, was necessary to train an effective and stable agent, although the system achieved a powerful performance. It can be found that a more explicit modeling or direct combination of coherence incentives of phrase transitions might also improve sentence-level coherence. Finally, the suggested DRL+MMR model shows a flexible and strong method for the query-based summarization task, able to learn effective policies for selecting sentences automatically. It successfully handles the main challenges of redundancy reduction, information coverage, summary coherence, and query relevance. Future work: Develop the system to be applied to a huge, diverse document corpus. Enhance the coherence system and apply developed DRL approaches like curriculum learning or transformer-based agents to further enhance the summarization quality. Also, for future work, combination query expansion assesses the DRL agent address the shortest queries effectively. Extending the reward function to contain more criteria.


## ACKNOWLEDGMENT

This research is supported by my respected supervisor Prof. Dr. Wafaa Mohammed Saeed, for giving me the motivation to fulfil the requirements of my PhD research in the University of Babylon College of Information Technology.


## REFERENCES


[1] Abualigah, L., Bashabsheh, M.Q., Alabool, H., Shehab, M. (2019). Text summarization: A brief review. In Recent Advances in NLP: The Case of Arabic Language, pp. 1-15. https://doi.org/10.1007/978-3-030-34614-0_1

[2] El-Kassas, W.S., Salama, C.R., Rafea, A.A., Mohamed, H.K. (2021). Automatic text summarization: A comprehensive survey. Expert Systems with Applications, 165: 113679. https://doi.org/10.1016/j.eswa.2020.113679

[3] Wafaa, A.H. (2024). Extracting key-phrase embedding using deep average network and maximal marginal relevance to enhance information retrieval. Journal of University of Babylon for Pure and Applied Sciences, 32(2): 80-91. https://doi.org/10.29196/jubpas.v32i2.5268

[4] Zhao, M., Yan, S., Liu, B., et al. (2021). QBSUM: A large-scale query-based document summarization dataset from real-world applications. Computer Speech & Language, 66: 101166. https://doi.org/10.1016/j.csl.2020.101166

[5] Molla, D. (2017). Towards the use of deep reinforcement learning with global policy for query-based extractive summarisation. arXiv preprint arXiv:1711.03859. https://doi.org/10.48550/arXiv.1711.03859

[6] Hyun, D., Wang, X., Park, C., Xie, X., Yu, H. (2022). Generating multiple-length summaries via reinforcement learning for unsupervised sentence summarization. arXiv preprint arXiv:2212.10843. https://doi.org/10.48550/arXiv.2212.10843

[7] Mohammed, M.B., Al-Hameed, W. (2021). New algorithm for clustering unlabeled big data. Indonesian Journal of Electrical Engineering and Computer Science, 24(2): 1054-1062. https://doi.org/10.11591/ijeecs.v24.i2.pp1054-1062

[8] Hasan, H.A., Al-Hussayni, K.H., Matloob, A.Z. (2025). Attention mapping for hallucination reduction in Arabic financial using retrieval-augmented generation systems. Journal of Intelligent Informatics, Networking, and Cybersecurity, 1(2): 2. https://doi.org/10.65445/3106-1192.1006

[9] Tan, F., Yan, P., Guan, X. (2017). Deep reinforcement learning: From Q-learning to deep Q-learning. In

International Conference on Neural Information Processing, pp. 475-483. https://doi.org/10.1007/978-3-319-70093-9_50

[10] Cajueiro, D.O., Nery, A.G., Tavares, I., et al. (2023). A comprehensive review of automatic text summarization techniques: Method, data, evaluation and coding. arXiv preprint arXiv:2301.03403. https://doi.org/10.48550/arXiv.2301.03403

[11] Saad, Y., Al Hameed, W. (2025). Toward salient key phrase for candidate topic detection. Kufa Journal of Engineering, 16(2): 215-233. https://doi.org/10.30572/2018/KJE/160213

[12] Rahman, N., Borah, B. (2015). A survey on existing extractive techniques for query-based text summarization. In 2015 International Symposium on Advanced Computing and Communication (ISACC), Silchar, India, pp. 98-102. https://doi.org/10.1109/ISACC.2015.7377323

[13] Zhang, W., Huang, J.H., Vakulenko, S., Xu, Y., Rajapakse, T., Kanoulas, E. (2024). Beyond relevant documents: A knowledge-intensive approach for query-focused summarization using large language models. In International Conference on Pattern Recognition, pp. 89-104. https://doi.org/10.1007/978-3-031-78495-8_6

[14] Bayatmakou, F., Mohebi, A., Ahmadi, A. (2022). An interactive query-based approach for summarizing scientific documents. Information Discovery and Delivery, 50(2): 176-191. https://doi.org/10.1108/IDD-10-2020-0124

[15] Yao, K., Zhang, L., Luo, T., Wu, Y. (2018). Deep reinforcement learning for extractive document summarization. Neurocomputing, 284: 52-62. https://doi.org/10.1016/j.neucom.2018.01.020

[16] Sotudeh, S., Goharian, N. (2023). Qontsum: On contrasting salient content for query-focused summarization. arXiv preprint arXiv:2307.07586. https://doi.org/10.48550/arXiv.2307.07586

[17] Lamsiyah, S., El Mahdaouy, A., Ouatik El Alaoui, S., Espinasse, B. (2023). 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, 14(3): 1401-1418. https://doi.org/10.1007/s12652-021-03165-1

[18] Liu, Y., Wang, Z., Yuan, R. (2024). QuerySum: A multi-document query-focused summarization dataset augmented with similar query clusters. Proceedings of the AAAI Conference on Artificial Intelligence, 38(17): 18725-18732. https://doi.org/10.1609/aaai.v38i17.29836

[19] Mridha, M.F., Lima, A.A., Nur, K., Das, S.C., Hasan, M., Kabir, M.M. (2021). A survey of automatic text summarization: Progress, process and challenges. IEEE Access, 9: 156043-156070. https://doi.org/10.1109/ACCESS.2021.3129786