IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Can Anaphora Resolution Improve Extractive Query-Focused Multi-Document Summarization?

SALIMA LAMSIYAH[1], ABDELKADER EL MAHDAOUY[2], AND CHRISTOPH SCHOMMER [1]
[1]Department of Computer Science, Faculty of Science, Technology and Medicine, University of Luxembourg, Luxembourg
[2]Modeling, Simulation and Data Analysis (MSDA), Mohammed VI Polytechnic University, Morocco

Corresponding author: Salima Lamsiyah (e-mail: salima.lamsiyah@uni.lu)

ABSTRACT Query-Focused Multi-Document Summarization (QF-MDS) is the task of automatically generating a summary from a collection of documents that answers a specific user's query. Extractive methods are mainly based on identifying, selecting, and ranking sentences according to their relevance to the pre-given query. These methods have shown promising results, however, they may yield incoherent summaries, as pronominal anaphoric expressions may appear unbound. To deal with this issue, this paper proposes a novel method that leverages the potential of both contextual embeddings as well as anaphora resolution methods. More specifically, Sentence-BERT (SBERT) model is used o generate contextual embeddings for the document's sentences and the user's query. Additionally, the SpanBERT model is used to resolve unbound pronominal references in the input documents' sentences with the aim of improving the cohesiveness of the generated summaries. We conducted a comprehensive comparative analysis using quantitative and qualitative evaluations with other state-of-the-art systems on the standard DUC'2005 and DUC'2007 datasets. The obtained results have shown that the proposed method is competitive and performs better than recent query-focused multi-document summarization systems on some ROUGE evaluation measures. In addition, human evaluation results further verify that our method was able to generate more informative, cohesive, and less-redundant summaries.

INDEX TERMS Query-Focused Multi-Document Summarization, Contextual Embeddings, Anaphora Resolution, Sentence-BERT, SpanBERT.

I. INTRODUCTION

With the constant expansion of textual information on the web, there is an increasing demand for tools that facilitate users' access to pertinent information. In particular, Automatic Text Summarization (ATS) has been attracting widespread interest in the last few years. ATS is a research area in the context of Natural Language Processing (NLP) whose goal is to automatically process and synthesize texts while preserving their salient aspects. Essentially, ATS systems allow users to find the relevant information corresponding to their needs and help them save their information access time. Generally, ATS methods can be divided into two main branches: (i) *Extractive summarization*, which produces summaries by identifying and extracting the most relevant sentences from the source documents [1]. (ii) *Abstractive summarization*, which generates summaries by reformulating and fusing ideas and often by using a new lexicon [2]. The research addressed in this work focuses on the *extractive approach*, more specifically, on the Query-Focused Multi-Document Summarization (QF-MDS) task.

QF-MDS represents an effective tool to deal with the rapid growth of textual information, which aims to generate an informative and concise summary from a cluster of topic-related documents that answers a specific user's query [3]. However, introducing query-focused multi-document into the summarization task causes new difficulties and challenges: i) capture the semantic information of the documents' sentences and the users' s queries, which helps produce relevant summaries to the input queries; ii) deal with information redundancy, which presents a major issue in multi-document summarization; and iii) manage the problem of cohesion that is necessary to produce cohesive summaries.

Therefore, to address the issues mentioned above, we propose an extractive QF-MDS method that incorporates contextual embedding methods as well as coreference resolution techniques, which are necessary for any text summarization task. On the one hand, we use the recently developed pre-trained Sentence-BERT (SBERT) model [4] to capture the semantics of the documents' sentences and the users' queries. SBERT is a variant of the BERT model [5] that employs a siamese network architecture to represent variable-length sentences by dense vectors in a low-dimensional vector space, wherein semantically similar sentences are closer together. In fact, SBERT has achieved state-of-the-art performance in several NLP tasks, and it provides sentence embeddings that can capture the contextual information and structure of sentences, which can then be compared using the cosine similarity measure. On the other hand, even though extractive methods have shown promising performances, the generated summaries may contain incoherent sentences, as pronominal anaphoric chains may appear unbound [6]. In this work, we propose to solve this issue by incorporating an anaphora resolution component in the pipeline of our method, which resolves the broken pronominal anaphoric expressions in the input sentences aiming to improve the cohesiveness of the extractive summarization. For this purpose, we use a state-of-the-art system, namely the SpanBERT model [7]. SpanBERT uses a modified version of BERT's architecture [5] to capture the meaning of words and phrases in the context of a sentence. One of the key features of SpanBERT is its ability to perform anaphora resolution, which is the task of identifying the referents of pronouns and other noun phrases in a sentence. To the best of our knowledge, this is the first work that uses SpanBERT for anaphora resolution in the context of extractive query-focused multi-document summarization.

Furthermore, to address the problem of redundancy, several methods have been proposed, including heuristic post-processing such as counting new bi-grams [8] or dynamic scoring that compares each source sentence with the current summary like Maximal Marginal Relevance (MMR) [9]. However, most of these methods rely on lexical features without semantic representation learning. Thus, in our method, we incorporate SBERT embedding representations into the MMR method to re-rank the selected sentences aiming to produce query-relevant summaries that cover salient and non-redundant information. Moreover, the proposed QF-MDS method is unsupervised that does not require domain knowledge or labeled training data.

To summarize, the main contributions of this work are as follows:

1) Introduce an unsupervised extractive method for query-focused multi-document summarization that relies on contextual embeddings, namely the SBERT model, for sentence and query representation.
2) Leverage transfer learning from SpanBERT, a pre-trained language model fine-tuned on the CoNLL2011-2012 datasets, to handle the broken pronominal anaphoric expressions and improve the cohesion of the generated summaries.
3) Assess the performance of the proposed method on the standard DUC'2005 and DUC'2007 datasets using quantitative and qualitative methods based on ROUGE metrics [10] and human evaluations.

The remainder of this paper is structured as follows: Section II presents a brief account of the literature concerning extractive QF-MDS methods and coreference resolution in summarization systems. Section III describes the main steps of the proposed method, while Section IV presents and discusses the experimental results. Finally, the conclusion and some lines for future works are presented in Section V.

II. RELATED WORK

In this section, we first present the most prevalent state-of-the-art methods for the unsupervised and supervised extractive QF-MDS methods. Then, we provide a brief review of the literature on coreference resolution in text summarization methods.

A. UNSUPERVISED QF-MDS METHODS

Unsupervised extractive methods are mainly based on identifying and selecting sentences according to their relevance to the user's query. Several methods have been introduced for this task, which falls into different categories, including *graph-based*, *topic-based*, *optimization-based*, and *deep learning-based* approaches.

Generally, in **graph-based methods** [11]–[15], a graph is constructed, where the nodes are sentences of the documents and edges scores represent the correlation measure between these nodes. The query-dependent weights are then added to the edge score of each sentence and accumulated with the corresponding correlation score. In fact, several graph-based methods have been introduced for sentence scoring. In particular, a graph manifold ranking method is used to measure the relevance score of each sentence in the input documents according to the query [11]. In the same context, the wAASum system [12] uses a weighted archetypal analysis factorization method for sentence scoring. Recently, two novel graph-based methods have been introduced that use the fuzzy and transversal hypergraphs models to infer topic distributions of sentences [14], [15]. Furthermore, the **topic-based methods** [16]–[18] have also achieved encouraging results in extractive text summarization; they are mainly based on topic modeling methods, including the latent semantic analysis (LSA), the probabilistic LSA (pLSA), and the latent Dirichlet allocation (LDA). These methods derive an implicit representation of text semantics that describe the main topics of the original documents. Besides, the **optimization-based methods** have shown impressive results; they consider the task of query-focused summarization as an optimization problem where several methods have been proposed to solve this task. For instance, the SpOpt system [19] uses a sparse optimization with a decomposable convex objective function to extract the relevant sentences. Recently, the authors in [20], [21] have introduced CES and Dual-CES systems, respectively. The CES system uses the Cross-Entropy method [22] to select a subset of relevant sentences to the query, whose combination is predicted to produce a good summary. The Dual-CES employs a two-step dual-cascade optimization approach with saliency-based pseudo-feedback distillation to better handle the tradeoff between saliency and focus in the summarization task. Both CES and Dual-CES systems have obtained state-of-the-art performances on the three DUC'2005-2007 datasets.

Meanwhile, **deep learning-based methods** have gained much attention in the last few years [23]–[26]. In this context, the QODE system [25] uses the restricted Boltzmann machines and dynamic programming to generate query-relevant summaries. Additionally, other researchers have introduced a query-focused text summarization method that uses the stochastic Ensemble Noise Auto-Encoders to select the relevant sentences from an ensemble of noisy runs [26]. More recently, a novel method has been proposed for extractive QF-MDS based on sentence embedding, BM25 model, and MMR method [3]. Specifically, BM25 and semantic similarity are used to measure the relevance of each sentence in the cluster according to the pre-given query where the top-ranked sentences are selected. Then, the MMR method is applied to re-rank the selected sentences and generate the final summary.

In this work, we re-implement the method proposed by [3], however, instead of combining BM25 and semantic similarity to score sentences, we use Sentence-BERT model [4], based on siamese architecture and fine-tuning mechanism, for sentence retrieving and re-ranking. Moreover, we integrate a state-of-the-art coreference resolution system, namely the SpanBERT model [7] into the pipeline of our method to resolve anaphoric coreferences chains in the generated summaries. We show that our method based on the SBERT and SpanBERT models has achieved promising results and outperformed the method introduced by [3].

B. SUPERVISED QF-MDS METHODS

Supervised extractive methods use labeled training data to build a model that predicts the relevant sentences to the input query. In most cases, the task of query-focused summarization is modeled as a sentence classification or a regression problem that is solved using supervised machine learning algorithms. Earlier research works have mainly focused on traditional machine learning algorithms including, Hidden Markov and Bayesian statistical models, which have been used to extract the sentences and query features to estimate the sentences' saliency [27], [28]. Moreover, the **HybHSum** system [29] uses hierarchical Latent Dirichlet Allocation to extract latent characteristics for documents' sentences and users' queries, which is then used to train a regression model to predict sentence scores. In the same context, the support vector regression model is also exploited in [30] to rank and predict the relevant sentences to the input query.

In recent years, with the success of supervised deep learning models in various natural language processing tasks, including generic multi-document summarization [31], [32], several research works have exploited the benefit of these models to improve the query-focused multi-document summarization task [33]. On the one hand, convolutional neural networks (CNN) have been widely used in this context showing promising results. For instance, the **AttSum** system [8] uses a convolutional neural network with an attention mechanism to automatically learn the sentences and the document cluster embeddings, which are then used to jointly tackle query relevance and sentence saliency ranking tasks. Similarly, the **SRSum** (Sentence Relation-based summarization) system [34] applies CNNs with an attention mechanism to automatically learns useful latent features by jointly learning representations of query sentences, content sentences, and title sentences as well as their relations. In the same context, the **CRSum-SF** system [35] combines both convolutional and recurrent neural networks with an attention mechanism to automatically learns useful contextual features by jointly learning representations of sentences and similarity scores between a sentence and sentences in its context. Extensive experiments, conducted on the standard DUC datasets, have proven the effectiveness of the latter systems; they have achieved significant performances outperforming the traditional-based machine learning methods in terms of ROUGE scores [10].

On the other hand, query-focused summarization methods that are based on the Transformer architecture [36] have shown impressive results. The Transformer is mainly based on self-attention instead of recurrent layers in an encoder-decoder model, which has achieved state-of-the-art results in language understanding. In this context, a coarse-to-fine modeling framework has been developed for the QF-MDS task [37] that exploits the potential of the pre-trained BERT model [5]. In particular, the framework is composed of three main components; i) a *relevance estimator* to retrieve relevant passages to the query, b) *an evidence estimator* that uses BERT to isolate segments likely to contain answers to the query, and c) a *centrality estimator* that finally selects the sentences to include in the summary. The developed framework has shown to be robust across domains and query types. Moreover, other research works have used the transformer encoder-decoder to generate focused abstractive summaries [38], [39], which extend the baseline models with new components to encode the queries together with multiple documents in a hierarchical setting. The empirical results have demonstrated that the proposed methods bring substantial improvements over several strong baselines.

To summarize, supervised deep learning models have proven to be powerful for query-focused multi-document summarization, however, they demand high computational power and a large amount of labeled training data which are not always available. Hence, generalizing supervised deep learning methods to new domains and languages remains a challenging task.

C. COREFERENCE RESOLUTION IN TEXT SUMMARIZATION

Extractive-based text summarization methods aim to identify and extract the most relevant textual segments from source documents and assemble them in an adequate way to form the final summary. These methods have shown promising results, however, the generated summaries may contain incoherent sentences, as pronominal coreferences may appear unbound [6]. In fact, coreference resolution is a fundamental task in NLP that has a great impact on understanding the semantics of texts; it aims to find all references to the same entity in a document [40].

Nevertheless, a few text summarization methods have been proposed in the literature that takes into consideration the coreference resolution issue. The first introduced text summarization method was based on the idea that the longest coreference chain presents the main topic of the original document, while the shorter chains indicate the subtopics. The final summaries consist of only those sentences related to the longest chain, thus helping to maintain the coherence of the generated summaries [41]. Other researchers have introduced two solutions for resolving broken coreference resolution expressions into text summarization methods [42]. The first solution uses an LSA-based sentence extraction method, based on both lexical and anaphoric information, to improve the quality of the generated summaries. The second solution is to scan the generated summaries looking for broken anaphoric expressions and replacing them with their corresponding entities. Both methods were assessed on the DUC'2002 dataset and showed significantly better performance than the versions of the system not processing anaphoric information.

Furthermore, other works have investigated the influence of pronominal anaphora resolution on term-based summarization [43]. The underlying hypothesis was that by integrating an anaphora resolver into the term weighting process, it is possible to obtain more accurate frequency counts of concepts referred to by pronouns. The experimental results have demonstrated the effectiveness of this hypothesis; it has substantially improved the informativeness and the coherence of the final generated summaries. In the same context, the COHSUM system [44], a cohesive extraction-based single document method, computes the distribution of the coreferences in the source documents. The main idea is that the relevant sentences are those providing the most references to other sentences and that other sentences are referring to. Experimental results using the DUC'2002 dataset have proven the effectiveness of COHSUM system in generating more cohesive summaries.

Besides, the G-FLOW system [45] handles the coreference resolution issue in the extractive multi-document summarization. It is mainly based on a joint model for selection and ordering that balances coherence and salience. The obtained results have shown that the G-FLOW system generates dramatically better summaries than other state-of-the-art systems. More recently, an efficient method has been introduced for handling unbound pronominal anaphoric expressions in the extractive single document summarization [6]. Similarly to [42], the proposed solution has been applied using two different scenarios. The first one is performed at the post-processing stage, which aims to find and fix unbound anaphoric expressions present in the generated summaries. The second one resolves the unbound pronominal coreferences and generates an intermediate representation of the source documents at the pre-processing stage of the proposed method. Both solutions were evaluated on the single document summarization dataset, namely CNN corpus [46]. Quantitative and qualitative evaluations have shown very encouraging performances.

In contrast to the existing methods, we propose an extractive method that exploits a deep neural network coreference resolution model, namely the SpanBERT model [7], for query-focused multi-document summarization. To the best of our knowledge, this is the first work that exploits the anaphora resolution in multi-document summarization, in particular, the query-focused task.

III. A COHESIVE QUERY-FOCUSED MULTI-DOCUMENT SUMMARIZATION (CohQFMDS-Sum) METHOD

The task of an extractive QF-MDS system is to generate a relevant and non-redundant summary Sum from a cluster of textual documents D that answers a specific user's query Q, such that $Sum \subseteq D$, and the constraint L on the summary length is not reached. As stated earlier, extractive text summarization methods have shown promising performance. However, the generated summaries may lack cohesiveness since they sometimes contain broken pronouns. To handle this issue in the extractive summarization process and enhance the cohesiveness of the produced summaries, we utilize the SpanBERT model [7] along with some rule-based heuristics.

As depicted in Figure 1, the proposed system, named **CohQFMDS-Sum**, consists of the following components:

- **Text pre-processing** involves cleaning and preparing the input text for further analysis. It mainly focuses on representing the input documents by a set of sentences.
- **Anaphora resolution** aims to enrich the semantic information by replacing the broken pronouns in the obtained sentences with their corresponding antecedents or entities, thus reducing ambiguity and improving cohesion.
- **Sentence and query representation** leverages the potential of transfer learning from Sentence-BERT model [4] to map the input sentences and user's query into embedding vectors that capture their semantic meaning.
- **Sentence retrieval and re-ranking** retrieves the relevant sentences from the input documents based on their similarity to the query using the cosine similarity function (Eq. 1). Then, the top-k extracted sentences are re-ranked using the MMR method [9] (Eq. 2). This process aims to maximize the relevance of the selected sentences to the user's query while minimizing redundancy.
- **Post-processing** involves additional processing, including some rule-based heuristics to avoid redundancy in the final produced summaries, as well as using a sentence ordering method.

We successively provide a detailed description of each of these steps in the following subsections.

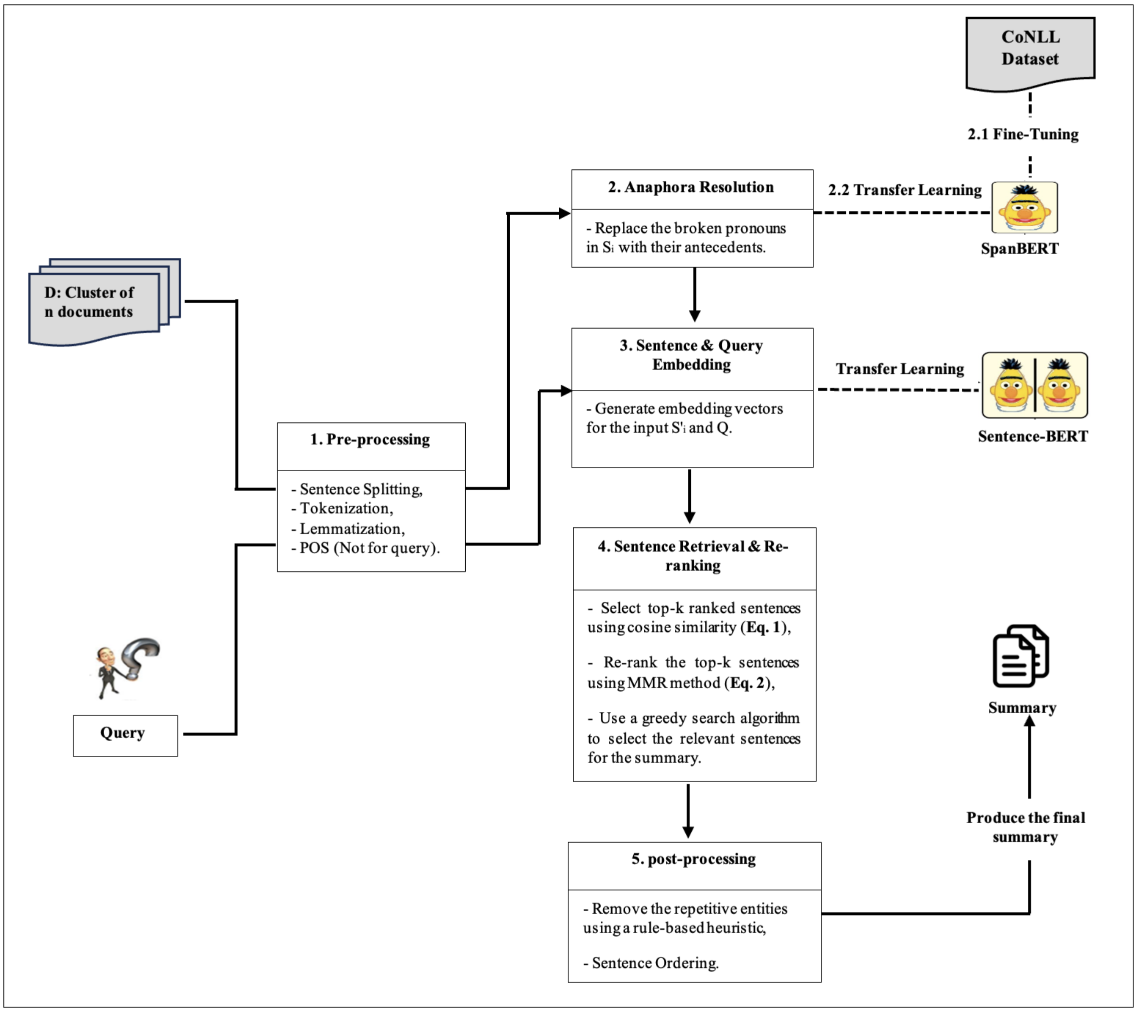
CoNLL Dataset
2.1 Fine-Tuning
2.2 Transfer Learning
SpanBERT
2. Anaphora Resolution
- Replace the broken pronouns in Sᵢ with their antecedents.
D: Cluster of n documents
3. Sentence & Query Embedding
- Generate embedding vectors for the input S'ᵢ and Q.
Transfer Learning
Sentence-BERT
1. Pre-processing
- Sentence Splitting,
- Tokenization,
- Lemmatization,
- POS (Not for query).
4. Sentence Retrieval & Re-ranking
- Select top-k ranked sentences using cosine similarity (Eq. 1),
- Re-rank the top-k sentences using MMR method (Eq. 2),
- Use a greedy search algorithm to select the relevant sentences for the summary.
Summary
Query
Produce the final summary
5. post-processing
- Remove the repetitive entities using a rule-based heuristic,
- Sentence Ordering.

FIGURE 1. Overall architecture of the proposed cohesive query-focused multi-document summarization system (CohQFMDS-Sum).

A. TEXT PRE-PROCESSING

In this step, we perform the morphosyntactic analysis of the input documents and the user's query following the commonly used preprocessing pipeline for text summarization tasks. More specifically, given a cluster of n documents, denoted as $D = \{d_1, d_2, ..., d_n\}$, we perform the following natural language processing subtasks:

1) **Sentence splitting:** for each input document d_i in the cluster D, we use the python library spaCy[1], in particular, the pre-trained model "en_core_web_md" to split d_i into a set of M sentences, denoted as $d = \{S_1, S_2, ..., S_M\}$. Then, we use regular expressions to remove special characters, such as redundant white spaces, XML/HTML tags, URLs, and email addresses.
2) **Tokenization:** for each sentence S_j in d_i, we first identify the individual words (tokens) within this sentence, and then we convert all these tokens into lowercase.
3) **Lemmatization:** we employ a lemmatizer from the NLTK[2] library to obtain the *canonical form* of each word w in S_j.
4) **Part-of-speech tagging (POS):** we use the NLTK POS tagger[3] to obtain the POS tag for each token w in S_j, which aims to identify its syntactic role in this sentence.
5) **Anaphora Resolution:** for each sentence S_j in d_i, we first verify if this sentence contains a broken anaphoric expression (pronoun) based on the POS tag of each word w in S_j. Then, we use a state-of-the-art coreference resolution system, namely the SpanBERT model [7] to resolve the broken pronominal anaphoric references in such a way as to eliminate coreference in the input sentences. Due to its paramount importance in the proposed method, the anaphora resolution step is further described in the next section.

Noticing that we performed tokenization, lowercasing,

[1] https://spacy.io/

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

[3] https://www.nltk.org/book/ch05.html

lemmatization, and special characters removal, using the spaCy library and regular expressions, to process the user's query and represent it as a simple sentence Q.

B. ANAPHORA RESOLUTION

Coreference resolution and anaphora resolution are related but distinct natural language processing tasks. Coreference resolution is the task of identifying all expressions in a text that refer to the same entity, while anaphora resolution is a specific type of coreference resolution that deals with the resolution of pronouns and their antecedents [40]. We focus on pronominal anaphora resolution, which is a crucial step for extractive text summarization tasks that helps reduce confusion and inaccuracies in the generated summaries. Most existing models in the literature have been proposed for coreference resolution [7], [47], [48]. In this work, however, we specifically focus on anaphora resolution within the context of extractive text summarization. To this end, we fine-tune the SpanBERT model for the anaphora resolution task on the CoNLL2011-2012 datasets [49], [50] - large datasets of generic texts that contain around 7,000 pronoun occurrences. Then, we use the fine-tuned SpanBERT model to find the antecedents of the broken pronouns in the input documents.

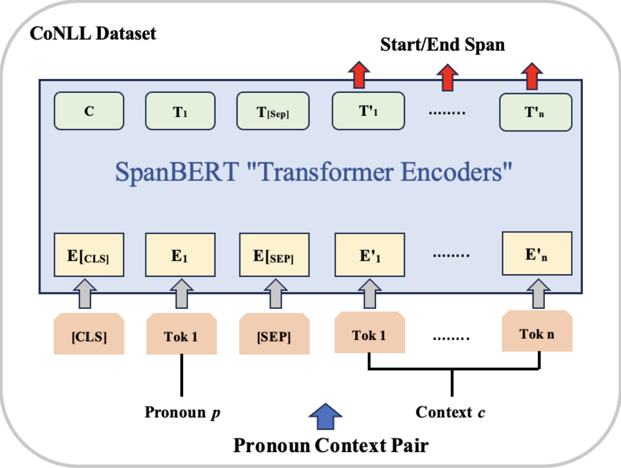

FIGURE 2. Process of SpanBERT Fine-tuning.

1) Fine-tuning SpanBERT model

The SpanBERT language model [7] is a variant of the BERT model [5] that has shown promising performance for the coreference resolution task outperforming other alternatives [7]. In contrast to BERT, the SpanBERT model is pre-trained to predict masked text spans rather than masked tokens. Furthermore, the SpanBERT model has shown to be more appropriate for tasks like anaphora resolution and question answering, where the desired output is a text span (e.g., a noun phrase) rather than just an individual noun [7], [51].

To fine-tune the SpanBERT model for antecedent learning (Step 2.1, Figure 1), we follow the recommendations in the literature for fine-tuning pre-trained language models [5], [7], with particular emphasis on fine-tuning SpanBERT for the anaphora resolution task [51]. This step aims to adjust the parameters of the general SpanBERT model by using the inputs and outputs of the CoNLL2011-2012 datasets [49], [50] for the anaphora resolution task. Therefore, we preprocess the latter datasets using the same NLP pipeline discussed in Section III-A. Thus, we create a list of pronouns (i.e., P) by selecting the words that the POS tagger marks as PRP (personal pronoun) or PRP$ (possessive pronoun). For each pronoun $p \in P$, we identify its context c which consists of the sentence that contains this pronoun and the preceding sentence. The SpanBERT model takes as input the tuples of the form $< c, p >$ as illustrated in Figure 3, which require to be tokenized and encoded into the same format that is used for training BERT. Thus, the input $< c, p >$ is first passed to BERT's tokenizer that adds two special tokens: **[CLS]** and **[SEP]** to help the model understand the input's structure. The first token **[CLS]** represents the classification output, while the second token **[SEP]** separates the context c from the pronoun p. Following this, we map the subword tokens to their corresponding integer IDs based on SpanBERT's vocabulary. Segment IDs are generated to distinguish the pronoun and the context parts of the input, allowing the model to learn the relationships between them. Finally, the encoded input is passed to a softmax layer to predict the text span which likely represents the antecedent a for each pronoun p based on its context c, formally defined as the $Probability(a|c, p)$ [51].

The model can identify multiple potential antecedents for a pronoun, each with a probability score indicating its likelihood of being the correct one. When an antecedent is predicted with a probability greater than 0.9 (based on our tuning), it is considered the resolved pronoun. For instance, as shown in Figure 3, the input to the SpanBERT model is $< c_1, p_1 >$ encoded as $[CLS]c_1\ [SEP]p_1$. The output of the model would be a tuple like $< s_1 =$ "The election"$, prob = 0.95\ >$. The text span s_1 would be the antecedent a of pronoun p_1 as it is identified with a probability greater than 0.9. It is worth mentioning that the SpanBERT model was specifically fine-tuned for anaphora resolution within the task of text summarization. However, the same model can be applied to tasks beyond summarization, such as question-answering and machine translation systems or any other context where anaphora resolution is required.

2) Anaphora Resolution with the Fine-tuned SpanBERT

Since we propose an unsupervised method for query-focused multi-document summarization, we did not fine-tune the SpanBERT on our text summarization datasets. However, we leverage the potential of transfer learning by directly utilizing the SpanBERT, which has already been fine-tuned on the CoNLL2011-2012 datasets [49], [50]. The fine-tuned model has demonstrated its effectiveness in predicting pronoun antecedents within their contextual context [51].

Formally, given an input document $d = \{S_1, S_2, ..., S_M\}$

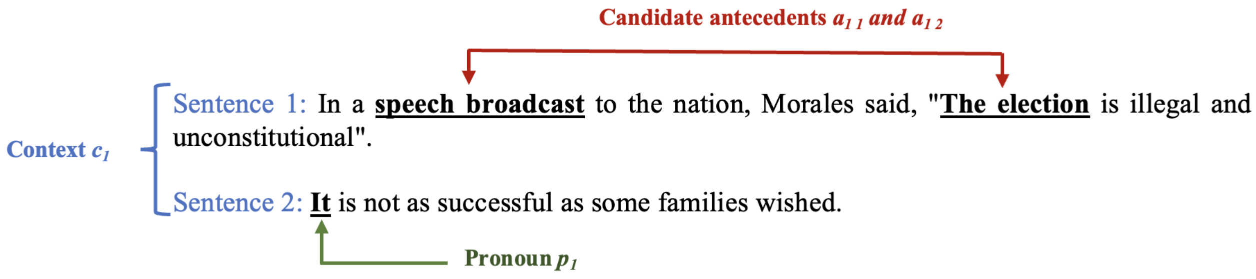
Candidate antecedents $a_{1\,1}$ and $a_{1\,2}$
Context c_1
Sentence 1: In a speech broadcast to the nation, Morales said, "The election is illegal and unconstitutional".
Sentence 2: It is not as successful as some families wished.
Pronoun p_1

FIGURE 3. Illustration of the Notation used For SpanBERT Fine-tuning.

that consists of M sentences, we first extract the pronouns contained in this document based on the part-of-speech tag of each word. Let $P = \{p_1, p_2, ..., p_m\}$ be the set of all pronouns in d in their order of appearance, we define the context c_k for each pronoun p_k as two consecutive sentences $c_k = (S_{j-1}, S_j)$, such that $2 \leq j \leq M$ and $1 \leq k \leq m$. Noticing that S_j is the sentence that contains the pronoun p_k, while for the pronoun p_k that occurs in S_1, its context c_k consists of only one sentence. Therefore, we employ the fine-tuned SpanBERT model as *a standalone solution* to predict the likely antecedent for the pronoun p_k from its context c_k, without the need for further training or fine-tuning. Then, we replace each pronoun with its corresponding antecedent in the input sentences. It is noteworthy that we have experimented with other coreference resolution systems, such as the NeuralCoref system[4]. However, the results have consistently shown that the fine-tuned SpanBERT performs better than this system.

As a result, we obtain intermediate documents $d_i^{'} = \left\{S_{1'}, S_2^{'}, ..., S_M^{'}\right\}$ where $1 \leq i \leq n$ and n is the number of documents in the cluster D, which have been enriched with solved coreferences. Indeed, our idea of resolving pronominal anaphoric expressions is justified by two main facts: 1) it is effective for obtaining unambiguous sentences, which helps improve the quality of the generated sentence embeddings. Especially, when a sentence contains many pronouns, the resulting embeddings may not accurately reflect the original meaning without enough context. 2) It helps produce more cohesive summaries without broken pronominal anaphoric chains.

C. SENTENCE AND QUERY REPRESENTATION

Sentence and query representation plays a significant role in extractive query-focused summarization methods. As previously mentioned, bag-of-words and word embedding representations are not able to fully capture the meaning of a sentence in one vector because they do not take into account the interactions between words or the order in which they appear. To address this issue and generate rich, semantically meaningful sentence embeddings, we leverage the potential of contextual embeddings using the current state-of-the-art Sentence-BERT model [4]. Specifically, there are two main approaches for leveraging transfer learning from Sentence-BERT embedding model, namely *feature-based* and *fine-tuning*. *Feature-based approach* uses the pre-trained SBERT model to extract fixed features for the input sentences, which can be used as input to the task at hand without any other modification. Besides, *fine-tuning approach* consists of re-training the pre-trained SBERT parameters on the down-stream task using task-specific data.

In our method, we use the SBERT embedding model as a feature extractor to generate rich semantic embedding vectors for the input query Q and for each sentence $S_j^{'}$ in $d_i^{'}$ such that $1 \leq j \leq M$ and $1 \leq i \leq n$, where n is the number of documents in the cluster D. The generated embedding vectors are denoted as $\overrightarrow{Q}$ and $\overrightarrow{S_j^{'}}$. Noticing that after the pre-processing and anaphora resolution steps, we represent the cluster of documents D by a set of sentences, denoted as $D = \left\{S_1^{'}, S_2^{'}, ..., S_N^{'}\right\}$ where N is the total number of sentences contained in the cluster D.

D. SENTENCE RETRIEVAL AND RE-RANKING

Sentence Retrieval and Re-ranking are essential steps in the extractive query-focused summarization process. In the *Sentence Retrieval* step, the system initially identifies a set of relevant sentences from a collection of documents that answer the user's query. Then, to further improve the quality of the generated summaries, *Sentence Re-ranking* is employed to reduce redundancy and maintain informativeness. The two steps are subsequently described as follows:

1) Sentence Retrieval

Let $D = \left\{S_1^{'}, S_2^{'}, ..., S_N^{'}\right\}$ be a cluster of textual documents consisting of N sentences, and let Q be a user's query. To measure the relevance score of each sentence $S_l^{'}$ in the cluster D according to the user's query Q, we use the cosine similarity as the semantic similarity metric, where $1 \leq l \leq N$. As defined in Equation 1, we compare the input query to each sentence in the cluster of documents by measuring the cosine similarity on their embedding vectors. To obtain these embeddings, both the sentences in the cluster and the query

[4] https://github.com/huggingface/neuralcoref

are transformed into dense vectors in a high-dimensional space using the Sentence-BERT embedding model (as described in Section III-C). By adopting this approach, we effectively capture the semantic information, enabling a more insightful and meaningful comparison between the query and the sentences within the cluster.

$$RelScore(S_l^{'}, Q) = cosSim(\overrightarrow{S_l^{'}}, \overrightarrow{Q}) = \frac{\overrightarrow{S_l^{'}} \cdot \overrightarrow{Q}}{||\overrightarrow{S_l^{'}}|| \cdot ||\overrightarrow{Q}||} \quad (1)$$

Where $RelScore$ is the relevance score of each sentence $S_l^{'}$ in D with respect to the query Q, $\overrightarrow{S_l^{'}}$ denotes the embedding vector of the sentence $S_l^{'}$ in the cluster D, and $\overrightarrow{Q}$ is the embedding vector of the input query Q. It is worth mentioning that we also used the negative Manhattan and negative Euclidean distances as similarity measures. However, the results were found to be almost identical. Therefore, based on the obtained relevance scores $RelScore(S_l^{'}, Q)$, we iteratively retrieve the top-k ranked sentences such as $k \in \{50, 100\}$ and consider them as candidates sentences for the final summary.

2) Sentence Re-ranking

Given the top-$k = \left\{S_1^{'}, S_2^{'}, ..., S_k^{'}\right\}$ retrieved sentences, we use a ranking algorithm to re-rank these sentences intending to produce Q-relevant and non-redundant summaries. As already mentioned, several methods have been proposed in the literature, including counting new bi-grams [8] and dynamic scoring using the Maximal Marginal Relevance (MMR) method [9]. However, these methods often lack semantic representation, relying mainly on lexical features. To this end, we employ a modified version of the MMR method [9], where we incorporate the sentence embeddings generated by the Sentence-BERT model [4].

The MMR method combines two main components: i) Relevance Score where each sentence is ranked based on its relevance to the query using some similarity measure, such as cosine similarity between the sentence and the query, and ii) Diversity Score that aims to ensure diversity in the summary by calculating a similarity score between each sentence and the sentences already selected for the summary. This score represents how similar the new sentence is to the sentences already in the summary. Therefore, as defined in Equation 2, for each sentence $S_p^{'}$ in the top-k selected sentences (where $1 \leq p \leq k$), we first calculate the relevance score of $S_p^{'}$ with respect to the input query Q using the cosine similarity (Eq. 1). Next, we compute its diversity score with the sentences already selected for the summary by also using the cosine similarity on their embedding vectors. Then, we combine linearly the relevance and the diversity scores to obtain the MMR score of the sentence $S_p^{'}$, denoted as $score_{MMR}(S_p^{'})$. Note that the sentence $S_p^{'}$ has a high marginal relevance score if it is both relevant to the query and contains minimal similarity to the previously selected sentences.

$$\begin{aligned} score_{MMR}(S_p^{'}) = \text{Argmax}_{S_p^{'} \in \text{top-}k \setminus Sum}[\lambda RelScore(S_p^{'}, Q) \\ - (1-\lambda) \max_{S_r \in Sum} DivScore(S_p^{'}, S_r^{'})] \\ 1 \leq p \leq k, p \neq r \end{aligned} \quad (2)$$

Where the $RelScore(S_p^{'}, Q)$ represents the relevance score of the sentence $S_p^{'}$ according to the input query Q (Eq. 1), $DivScore(S_p^{'}, S_r^{'})$ is the diversity score computed using the cosine similarity on the embedding vectors of the current sentence $S_p^{'}$ and the already selected sentences as summary $S_r^{'}$ (Eq. 1), top-k denotes the selected sentences obtained in the previous step, Sum subset of sentences in top-k already selected as a summary, and top-$k \setminus Sum$ represents the set of unselected sentences in top-k. Additionally, λ is an interpolation coefficient in the range $[0.5, 0.95]$ with constant steps of 0.05 that balances the trade-off between relevance and diversity. Finally, based on the obtained MMR sentences' scores, we select the sentences that will be included in the final summary, where a new sentence is added to the current summary if the constraint on the summary length limit L is not reached, and the semantic similarity between this sentence and the already selected summary sentences is below a threshold τ.

E. POST-PROCESSING

The use of the SpanBERT system allows us to select sentences with resolved broken pronominal anaphoric expressions. However, as shown in the following example, the simple strategy of replacing every pronoun may cause redundant information and repetitive entity references in the final produced summaries.

- **Example $\mathbf{S_1}$: Morris Dees**, the co-founder of the Southern Poverty Law Center in Montgomery, Ala., and one of the attorneys for the plaintiffs, said **he (Morris Dees)** intended to enforce the judgment, taking everything the Aryan Nations owns, including its trademark name.

To address this issue, we apply a rule-based heuristic that for each sentence in the generated summary, it keeps the pronoun if it appears after its referents; otherwise, the pronoun is unbound and must be replaced by its entity. The main idea is to substitute only the pronominal anaphoric expressions p_k whose contexts c_k are not present in the generated summary.

To further improve the informativeness and cohesiveness of the generated summaries, we perform ***sentence reordering***, which refers to sorting the selected sentences into the appropriate coherent order. For this purpose, we use the Chronological Ordering algorithm [52] that rearranges the sentences for each query based on the time stamp and the position in the source document.

IV. EXPERIMENTAL RESULTS

In this section, we first present the used datasets, the evaluation metrics, and the experimental setup. Then, we discuss the obtained results, intending to verify the following hypotheses:

- *Hypothesis 1*: The Bi-Encoder Sentence-BERT model produces better sentence embeddings than the Cross-Encoders BERT and the SpanBERT models [5], [7].
- *Hypothesis 2*: Resolving pronominal anaphoric chains, using the SpanBERT [7] system improves the cohesiveness of the generated summaries.
- *Hypothesis 3*: The proposed method is effective as compared to recent supervised and unsupervised extractive query-focused multi-document summarization methods.

A. DATASETS

The experiments use the standard DUC'2005-2007 benchmarks created by NIST[5] and considered the widely used corpora for evaluating the performance of query-focused summarization methods. As presented in Table 1, each dataset consists of a set of clusters, with each cluster having a single query and composed of an average of 25 English news articles. The gold standard summaries are provided by different experts, and the length of each summary is limited to 250 words, as required in DUC evaluations. More precisely, the DUC'2006-2007 datasets feature 4 expert-written summaries per cluster, whereas the DUC'2005 dataset has 4-9 human-written summaries per cluster. Additionally, each query contains the main topic followed by additional questions indicating the aspects that should the summarization system cover; e.g.:

"Same-sex schools. What are the advantages and disadvantages of same-sex schools?"

Furthermore, the Sentence-BERT model has been trained on the Natural Language Inference (NLI) dataset that is constructed by combining two datasets, including the Stanford Natural Language Inference (SNLI) [53] and the Multi-Genre NLI [54]. The SNLI contains 570.000 pairs of sentences, annotated with the labels *entailment*, *contradiction*, and *neutral*. The Multi-Genre NLI is a collection of 430.000 sentence pairs and covers various genres of spoken and written texts. Additionally, it was further fine-tuned on the Semantic Textual Similarity benchmark (STSb) [55], a popular dataset for evaluating the supervised STS systems where the task is to predict the semantic similarity score between a pair of two sentences using a regression objective function. The STSb benchmark includes 8628 pairs of sentences divided into three categories, including *captions*, *news*, and *forums*.

B. EXPERIMENTAL SETUP

Our current pipeline is based on a set of Python tools, including the TrecTools[6] library and the available implementation of SBERT[7] and SpanBERT[8] models. The preprocessing pipeline is implemented using SpaCy[9] and NLTK[10] libraries. We have used an Intel(R) Xeon(TM) CPU 3.00 GHz server equipped with Nvidia Tesla K40c GPU having 12 GB of RAM to run all the experiments except the SpanBERT fine-tuning. We have tested two variants of the SBERT model: SBERT$_{BASE}$ and SBERT$_{LARGE}$. The SBERT$_{BASE}$ is designed to embed a sentence into 768-dimensional vectors, while SBERT$_{LARGE}$ provides sentence embedding vectors of 1024 dimensions. We reported the results of the SBERT$_{BASE}$ model because their results remained roughly the same. Additionally, we have used SpanBERT$_{BASE}$ model fine-tuned on the CoNLL2011-2012 datasets [49], [50] for 20 epochs with 2e-5 learning rate and 32 batch size, where the fine-tuning has been done on the Iris cluster[11] at the University of Luxembourg, which features 96 Nvidia V100 GPU-AI accelerators with Skylake or Broadwell processors. Specifically, we used 4 GPUs with ten cores and one node.

Furthermore, we have used three hyperparameters, including the number of top-ranked sentences k, the interpolation coefficient λ, and the threshold τ. Such that $k \in \{50, 100\}$, while λ and τ are in range $[0.5, 0.95]$ with constant steps of 0.05. To optimize these hyperparameters, we shuffle and randomly sample 20 clusters from the DUC'2006 dataset to create a small held-out set. Then, a grid search is performed on the held-out set that gave us a total of 200 feasible combinations. Accordingly, the optimized values of λ, τ, and k are 0.9, 0.85, and 50, respectively. Moreover, the sentences with the highest scores are selected to compose the summary, where the total of the selected sentences depends on the defined compression rate. As the golden standard summaries of DUC'2005-2007 datasets comprise about 250 words, the same compression rate was used in all our experiments. Besides, for the statistical significance test, we have applied the *paired t-test* [56] to determine whether there is a significant difference in performance among all the evaluated models. We have attached a superscript to the performance number in the tables when the $p - value < 0.05$.

C. EVALUATION METHODS

To assess the effectiveness of the proposed method, we have used ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [10] and Human Evaluation for quantitative and qualitative evaluations, respectively.

For the ***quantitative evaluation***, we have used ROUGE-N (ROUGE-1 and ROUGE-2) and ROUGE-SU4. ROUGE-N determines the similarity between the systems summaries and a set of gold summaries based on the n-gram overlap, whereas ROUGE-SU4 determines the overlap of skip-bigram between a system summary and a collection of reference summaries with a max distance of four words. We have reported the obtained recall performance of ROUGE-1 (R-

[5] https://duc.nist.gov/
[6] https://pypi.org/project/trectools/
[7] https://www.sbert.net/
[8] https://huggingface.co/SpanBERT/spanbert-base-cased/tree/main
[9] https://spacy.io/
[10] https://www.nltk.org/
[11] https://hpc-docs.uni.lu/systems/iris/

TABLE 1. Statistics of DUC'2005-2007 Datasets. *Num docs* is the number of documents in each cluster. *Sum length* indicates the number of words in gold summaries. *Num gold sum* is to the number of human summaries written for each cluster.

Datasets	Clusters	Num docs	Sentences	Queries	Sum length	Num gold sum	Data source
DUC'2005	50	32	45931	50	250	4	TREC
DUC'2006	50	25	34560	50	250	4	AQUAINT
DUC'2007	45	25	24282	45	250	4-9	AQUAINT

1), ROUGE-2 (R-2), and ROUGE-SU4 (R-SU4) using the official ROUGE toolkit (version 1.5.5) with standard options settings[12] used for assessing extractive QF-MDS systems. Furthermore, it is worth mentioning that the ROUGE method focuses on the informativeness of the produced summary; a recent research work [57] has demonstrated that no other automatic metric consistently achieves better performance than the ROUGE method in evaluating text summarization systems.

Besides, ***qualitative evaluation*** represents a challenging task for automatic text summarization, especially for multi-document summarization. *Human Evaluation* is a subjective task that requires a deep understanding of the original texts where the same person could write very different summaries in a few weeks. Additionally, evaluating properties such as relevance, coherence, cohesion, readability, or co-reference resolution depends on several aspects such as background knowledge, or even linguistic skills. Thus, human evaluation is very costly and time-consuming, especially when evaluating multi-document summarization systems, because the size of the input documents makes the evaluation even more complex. Although generating a summary is a difficult task in itself, assessing the quality of the generated summaries is another matter altogether. For evaluating our method, we follow the previous works [6], [37]–[39] where the generated summaries are evaluated in a judgment elicitation study via Amazon Mechanical Turk[13]. More precisely, we randomly generate samples from DUC'2005 and DUC'2007 datasets, and each sample is evaluated by English native speakers from USA and UK. The turkers are asked to rate query-summary pairs based on three aspects: *a) Succinctness*: Does the summary deal with redundant and unnecessary information? *b) Cohesion*: Does the summary contain coherent sentences and make logical sense? and *c) Relevance to the query*: Does the summary answer the query? *Succinctness* and *Cohesion* were rated using a five-point Likert scale, while for the *Relevance*, the participants were asked to read the summary and decide for each sentence whether it is query-relevant, query-irrelevant, and partially relevant. Relevant sentences were awarded a score of 5, partially relevant ones a score of 2.5, and 0 otherwise. Sentence scores were averaged to obtain a relevance score for the whole summary. The obtained results are summarized in Table 4 and discussed in section IV-E.

D. EFFECTIVENESS OF SBERT CONTEXTUAL EMBEDDING MODEL

Several experiments were conducted to evaluate the performance of the SBERT model for the unsupervised extractive query-focused multi-document summarization task (examine ***Hypothesis 1***). To this end, we have implemented the proposed method using three different text representation methods: word embeddings based on the average of GloVe embeddings [58], the average of BERT and SpanBERT embeddings [5], [7], and SBERT embeddings [4]. The obtained results of these methods (denoted as **GloVe-Sum**, **BERT-Sum**, **SpanBERT-Sum**, or **CohQFMDS-SBERT-Sum** according to the embedding model that is used) are summarized in Table 2.

As shown in Table 2, the proposed method CohQFMDS-SBERT-Sum based on the pre-trained bi-encoder SBERT outperformed the GloVe-Sum, BERT-Sum, and SpanBERT-Sum models on most evaluation measures for the DUC'2005-2007 datasets. Specifically, based on the R-1 measure, the average performance of our method increased by approximately 2 percentage points compared to GloVe-Sum on the two datasets. Additionally, it can be seen from Tables 1 and 2 that the SpanBERT-Sum performed better than GloVe-Sum and significantly exceeded the performance of BERT-Sum. Additionally, the results show that directly using the output of the BERT model by averaging BERT embeddings leads to rather poor performance, which is worse than computing the average of GloVe embeddings. This can be attributed to the fact that the BERT model is trained on a masked language model, where the output vectors are tied to individual tokens rather than sentences, whereas summarization methods work with sentence-level representations. Furthermore, this finding aligns with previous studies [4], [59], which have demonstrated that BERT embeddings are not appropriate for unsupervised natural language processing tasks. Therefore, these noteworthy results confirm that utilizing the SBERT embedding model, which employs a siamese network structure and fine-tuning mechanism to capture semantics, can significantly enhance the performance of extractive QF-MDS systems when compared to other models such as GloVe, BERT, or SpanBERT embeddings. It is noteworthy that SBERT is trained on the NLI dataset, which is considered one of the largest and high-quality labeled corpus for textual entailment tasks. As a result, it helps the summarizer in selecting the most relevant information from the input documents that are logically entailed by the input query.

[12] -a -c 95 -m -n 2 -2 4 -u -p 0.5 -l 250

[13] https://www.mturk.com/

TABLE 2. ROUGE recall scores of GloVe-Sum, BERT-Sum, SpanBERT, and CohQFMDS-SBERT-Sum methods on DUC'2005-2007 datasets using ***SpanBERT-based solution***. for indicating statistical significance performances, the superscripts $number$ denotes significant improvement ($p -$ value < 0.05) over the method that has the same superscript $number$ attached.

	DUC'2005			DUC'2007		
	R-1	R-2	R-SU4	R-1	R-2	R-SU4
Avg. GloVe-Sum^{1}	38.84^{2}	8.28^{2}	14.34^{2}	42.63^{2}	10.83^{2}	16.46^{2}
Avg. BERT-Sum^{2}	37.26	7.56	12.84	39.25	8.22	14.32
Avg. SpanBERT-Sum^{3}	39.14^{2}	8.46^{2}	14.56^{2}	43.35^{1-2}	11.24^{2}	16.82^{2}
CohQFMDS-SBERT-Sum^{4}	41.87^{1-3}	9.74^{1-3}	16.2^{1-3}	45.76^{1-3}	12.64^{1-3}	18.85^{1-3}

E. EFFECTIVENESS OF ANAPHORA RESOLUTION

The main goal of these experiments is to address ***Hypothesis 2***: *Can the pronominal anaphoric resolution improve the cohesion of the generated summaries when using an extractive system?* We conducted several experiments on DUC'2005 and DUC'2007 datasets using both quantitative and qualitative metrics based on the ROUGE method and human evaluations, respectively. For quantitative evaluation, we evaluated our method using two different scenarios, described as follows:

- *Scenario 1*: We remove the anaphora resolution step from our pipeline, thus the final summaries are generated without resolving anaphora.
- *Scenario 2*: The final summaries are generated after resolving pronominal anaphoric expressions.

The obtained ROUGE recall scores are summarized in Table 3, while results from the *human evaluation* are presented in Table 4. A paired t-test [56] was performed between the ROUGE scores and a superscript is attached to the performance number in the table when the $p -$ value < 0.05. Moreover, a concrete example of the output of our system with the gold summary is illustrated in Table 6.

TABLE 3. ROUGE recall scores of Scenario 1 and Scenario 2 on DUC'2005 and DUC'2007 datasets of our method using the **SpanBERT-based solution**. The symbol † denotes statistical significant improvement ($p -$ value < 0.05) of Scenario 2 over Scenario 1.

	DUC'2005			**DUC'2007**		
	R-1	**R-2**	**R-SU4**	**R-1**	**R-2**	**R-SU4**
Scenario 1	40.17	8.68	15.75	43.75	11.57	17.96
Scenario 2	41.87†	9.74†	16.2†	45.76†	12.64†	18.85†

From Table 3, we observe that on the two used datasets, Scenario 2, using the *SpanBERT* anaphora resolution system, has achieved better performance and led to statistically significant improvements over Scenario 1 for almost all evaluation measures (R-1, R-2, R-SU4). For instance, an improvement of 2.01%, 1.07%, and 0.89% for R-1, R-2, and R-SU4 metrics was achieved on the DUC'2007 dataset compared to Scenario 1. These noteworthy results verify the effectiveness of the proposed solution for resolving anaphoric expressions.

TABLE 4. Human Evaluation on DUC'2005 and DUC'2007 benchmarks. **R**elevance, **S**uccinctness, **C**oherence ratings; **All** is the average of all the ratings; Highest score is shown in bold.

Systems	**R**el	**S**uc	**C**oh	All
LEAD-3	3.72	3.51	4.07	3.76
USE-Transformer-Sum [3]	4.13	3.96	3.92	4.0
CohQFMDS-SBERT-Sum	4.28	3.98	4.23	4.16
Gold Summary	4.31	4.01	4.36	4.22

Furthermore, for qualitative evaluation, we randomly sampled 20 query-cluster pairs from DUC'2005 and DUC'2007 (10 from each dataset). Then, we collected three responses (Relevance, Succinctness, and Cohesion) per query-summary pair. We compared the summaries created by our method CohQFMDS-SBERT-Sum, the USE-Transformer-Sum [3], the LEAD-3 baseline, and the GOLD summary (ground-truth upper bound). Table 4 presents the ratings of each system on DUC benchmarks. As can be seen, participants find that CohQFMDS-SBERT-Sum is more relevant with less redundant information compared to the LEAD-3 baseline. Moreover, it has shown comparable results to the USE-Transformer-Sum system in terms of relevance and succinctness scores; they are both based on contextual embeddings and MMR method [9] for sentence scoring and reranking. Finally, in terms of cohesion, our method produces more coherent summaries than the USE-Transformer-Sum and LEAD-3 systems achieving comparable results to the gold summaries. This further demonstrates the robustness of the proposed method in handling broken anaphoric expressions in the generated summaries.

F. COMPARISON WITH STATE-OF-THE-ART METHODS

Finally, to address ***Hypothesis 3***, we compare our method with the **best performing** recent state-of-the-art QF-MDS methods on the standard DUC'2005 and DUC'2007 datasets. Note that we report the results of the best-performing variant of our method **CohQFMDS-SBERT-Sum** that uses *SBERT* model for generating contextual embeddings and the *SpanBERT* model for anaphora resolution, while for the state-of-the-art systems, we report the results depicted in their corresponding papers. The overall ROUGE recall scores are summarized in Table 5.

The first set of analysis is performed to compare our method, **CohQFMDS-SBERT-Sum**, with the best-performing extractive **unsupervised** query-focused multi-document systems, including CES [20], Dual-CES [21], and USE-Transformer-Sum [3]. As depicted in Table 5, on the DUC'2005 dataset, our method has outperformed all the other methods including the best-performing Dual-CES system for all the evaluation measures. Additionally, on the DUC'2007 dataset and in terms of R-1 and R-2, our method has yielded better performance than the USE-Transformer-Sum and CES systems, while it has achieved comparable performance to the Dual-CES system. This can be because the Dual-CES system better handles the trade-off saliency and focus on the summarization process. Nevertheless, regarding the R-SU4 measure, our method has achieved the best performances; it has outperformed all the systems that we compared with on the two used DUC'2005-2007 datasets. Furthermore, the obtained results have shown that our method CohQFMDS-SBERT-Sum has outperformed the USE-Transformer-Sum method on the two DUC'2005-2007 datasets for all the evaluation measures. It's worth mentioning that the USE-Transformer-Sum combines the BM25 model [60] and the semantic similarity to select the relevant sentences to the input query while in our method we use only the semantic similarity, thus further validating the effectiveness of the SBERT embedding model.

The second set of analysis is conducted to compare the proposed method with recent **supervised** extractive QF-MDS methods namely, **HybHSum** [29], **AttSum** [8], **SRSum** [34], and **CRSum-SF** [35] systems (described in the related work section II-B). The **HybHSum** system is based on a probabilistic topic model for pattern discovery and a regression model for sentence score prediction. The **AttSum** and **SRSum** systems are based on convolutional neural networks with attention mechanisms. The **CRSum-SF** system is based on both convolutional neural networks and recurrent neural networks. ROUGE recall scores of these systems are presented in the second block of Table 5. As presented in Table 5, our method has outperformed all other systems in terms of R-1 and R-2 evaluation measures on the DUC'2005 dataset. Specifically, it has performed better than SRSum and CRSum-SF systems in terms of R-1 and R-2 evaluation measures, while it has achieved comparable performance to them on the DUC'2007 dataset. Furthermore, our method has shown comparable performance to the HybHSum system that has yielded the best R-1 score on the DUC'2007 dataset outperforming all the other methods. However, in terms of the R-2 score, our method has achieved far better performance than it. Lastly, in terms of the R-SU4 measure, our method has shown far better performances than all these methods on the two used datasets.

The overall comparison results show that our method has achieved better performances than the best-performing unsupervised state-of-the-art methods (CES, Dual-CES, and USE-Transformer-Sum) on DUC'2005-2007 datasets for most evaluation measures. Additionally, it shows promising results when compared to recent supervised deep learning-based state-of-the-art methods (SRSum, CRSum-SF). These findings demonstrate the effectiveness of our proposed method, which is based on the Sentence-BERT model and SpanBERT coreference resolution system, in generating relevant and coherent summaries.

G. ERROR ANALYSIS

To deepen the results, we conduct an error analysis that discusses the successes and failures of the proposed method. Table 7 shows examples of the generated summaries with fixed broken anaphoric expressions using our method CohQFMDS-SBERT-Sum. Table 6 shows some fragments of the generated summaries with such cohesion problems.

From Table 7, we observe that the produced summaries based on Scenario 1 contain sentences ($\mathbf{S}_2$, $\mathbf{S}_3$, $\mathbf{S}_6$, and $\mathbf{S}_7$) with broken pronominal anaphoric expressions, which negatively affect the cohesiveness and fluency of the generated summaries. However, our method SBERT-MT-Sum, based on the SpanBERT system and a rule-based heuristic, was able to handle this issue by replacing each broken pronoun with its corresponding entity (Scenario 2). It is worth noting that the proposed method performed the correct coreference substitution and thus improves the text quality (cohesion, fluency, and readability) of the final extractive generated summaries. Furthermore, as seen in Table 7, the produced summaries include sentences that are relevant to the input query. This makes sense since the SBERT model, based on the siamese architecture, can generate contextual embeddings that capture the meaning of the document sentences and input queries. Additionally, the generated summaries do not contain repetitive sentences, indicating that using SBERT embeddings with the MMR method can efficiently address the issue of redundancy. This is particularly important in the context of multi-document summarization, where the risk of selecting redundant sentences is higher than in single-document summarization.

Even though the proposed method produces satisfactory performances for extractive query-focused multi-document summarization, there is still scope for improvement. Hence, we investigate cases further where it goes wrong. For instance, from Table 6, we observe that the entity (Angelina Jolie) in sentence S_1 is repeated in the summary because the proposed method was not able to identify that the word "Jolie" and "Angelina Jolie" refer to the same entity; it pointed out that the pronoun "**she**" was unbound. Thus, a possible solution to this problem is to treat nominal coreferences, making the proposed method able to identify the word

TABLE 5. QF-MDS systems performance comparison on DUC'2005 and DUC'2007 datasets, using ROUGE recall scores (R-1, R-2, and R-SU4). The symbol "— —" indicates that the results are not available in their respective works. The highest scores (R-1, R-2, and R-SU4) for each group system are printed in boldface. The symbol ⋆ indicates the best-performing system for each measure.

	DUC'2005			DUC'2007		
	R-1	R-2	R-SU4	R-1	R-2	R-SU4
CES [20]	40.33	7.94	13.89	45.43	12.03	17.5
Dual-CES [21]	**40.82**	**8.07**	14.13	**46.02**⋆	**12.53**	17.91
USE-Transformer-Sum [3]	39.79	8.27	**15.52**	43.54	11.42	**18.54**
HybHSum [29]	— —	— —	— —	**45.6**	11.4	17.2
AttSum [8]	37.01	6.99	— —	43.92	11.55	— —
CRSum-SF [35]	39.52	8.41	— —	44.6	12.48	— —
SRSum [34]	**39.83**	**8.57**	— —	45.01	**12.8**⋆	— —
CohQFMDS-SBERT-Sum	**41.87**⋆	**9.74**⋆	**16.2**⋆	**45.76**	**12.64**	**18.85**⋆

TABLE 6. Examples of generated sentences with text quality problems from the DUC'2007 dataset.

- S$_1$ Jolie has worked constantly for the past 18 months and when she isn't on a set, **she (Angelina Jolie)** is usually reading or writing or acting on the stage.
- S$_2$ Quotas have the effect of marginalizing women, **she** argues; an objection **she (Sally Hamwee)** also raises against Labour's proposed ministry for women's affairs.
- S$_3$ We must immediately put into action practical water conservation measures which will reduce our water consumption and allows us to recycle water," **he (President Fidel Ramos' call to respond to El Nino)** said.
- S$_4$ As a result, doctors are seeing more patients whose regimens include treatments such as acupuncture and Chinese herbs, which Soto also takes. **They (Patients)** have also explained why acupuncture can kill pain, with extensive publicity helping make acupuncture popular in the world.

"Jolie" referred to as the entity "Angelina Jolie". Although the proposed method has indicated the correct referents, the substitution has not been necessary. Furthermore, for the sentence S_2, the proposed method did not identify the first pronoun and replace the second one in the same coreference chain. Moreover, we noticed that sentence S_3 includes unnecessary information that negatively impacts the quality of the generated summary. Additionally, sentence S_4 highlights an ambiguity detection of the pronoun "**They**", which refers to the entity "**doctors**" and not "**patients**". To address this issue, we plan to incorporate an ambiguity detection module in the anaphora resolution process to revise pronouns that may cause confusion.

V. CONCLUSION

In this paper, we proposed an effective unsupervised extractive method for query-focused multi-document summarization based on contextual sentence embeddings and anaphora resolution. The proposed method aims to produce a cohesive summary from a collection of documents that answers a specific user's query. Specifically, the main contributions of this work are summarized as follows: a) leverage the potential of the Sentence-BERT model to represent the documents' sentences and the input queries, b) improve the cohesiveness of the generated summaries by resolving the broken pronominal anaphoric expressions, and c) assess the robustness of the proposed method against recent supervised and unsupervised QF-MDS methods.

We conducted extensive experiments on the standard DUC'2005 and DUC'2007 datasets to examine the effectiveness of the proposed contributions. Our primary objective was to investigate whether anaphora resolution could improve the performance of extractive query-focused multi-document summarization. Through a combination of quantitative and qualitative evaluations, we were able to observe that integrating the anaphora resolution component into our pipeline had a significant impact on the cohesiveness of the generated summaries. Our work demonstrated that resolving broken anaphoric expressions is crucial in producing high-quality summaries that convey information accurately. Additionally, the SpanBERT model showed to be effective to address this issue and improve the quality and cohesiveness of extractive text summarization systems. Furthermore, the experimental results indicated that using the Sentence-BERT model for sentence and query embeddings achieved better performance than other models such as GloVe, BERT, SpanBERT, and the USE-Transformer. We also observed that the summaries generated using our approach were more relevant to the input queries and contained less redundant information. Additionally, our method demonstrated promising performance across different datasets (DUC'2005 and DUC'2007) when compared to the best-performing systems, including recent supervised deep learning-based methods like the SRSum system. It is worth noting that our proposed method is unsupervised and does not require labeled training data or domain knowledge. The overall results underscore the effectiveness of utilizing the Sentence-BERT model in conjunction with an anaphora resolution method based on the SpanBERT model for extractive query-focused multi-document summarization.

In the future, we envisage integrating an ambiguity detection module in the anaphora resolution process that revises those pronouns that can lead to misunderstandings. Moreover, we also plan to explore the performance of recent pretrained language models such as T5 (Text-To-Text Transfer Transformer) [61] to generate abstractive query-focused multi-document summaries. The abstractive approach produces summaries by concisely paraphrasing the document's content, which directly improves the cohesion and coherence of the generated summaries. Furthermore, ROUGE metrics [10] are mainly based on surface lexical similarities, and hence it is still challenging to accurately measure the similarity between a generated summary and the golden ones semantically. Therefore, we would like to investigate how recent sentences embeddings models can improve the ROUGE method.

REFERENCES

[1] Muhammad F Mridha, Aklima Akter Lima, Kamruddin Nur, Sujoy Chandra Das, Mahmud Hasan, and Muhammad Mohsin Kabir. A survey of automatic text summarization: Progress, process and challenges. IEEE Access, 9:156043–156070, 2021.

[2] Jingwei Cheng, Fu Zhang, and Xuyang Guo. A syntax-augmented and headline-aware neural text summarization method. IEEE Access, 8:218360–218371, 2020.

[3] Salima Lamsiyah, Abdelkader El Mahdaouy, Said Ouatik El Alaoui, and Bernard Espinasse. Unsupervised query-focused multi-document summarization based on transfer learning from sentence embedding models, bm25 model, and maximal marginal relevance criterion. Journal of Ambient Intelligence and Humanized Computing, pages 1–18, 2021.

[4] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China, 2019.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, 2019.

[6] Jamilson Antunes, Rafael Dueire Lins, Rinaldo Lima, Hilario Oliveira, Marcelo Riss, and Steven J Simske. Automatic cohesive summarization with pronominal anaphora resolution. Computer Speech & Language, 52:141–164, 2018.

[7] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 8:64–77, 2020.

[8] Ziqiang Cao, Wenjie Li, Sujian Li, Furu Wei, and Yanran Li. AttSum: Joint learning of focusing and summarization with neural attention. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 547–556, 2016.

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

[10] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain, 2004.

[11] Xiaojun Wan and Jianguo Xiao. Graph-based multi-modality learning for topic-focused multi-document summarization. In Twenty-First International Joint Conference on Artificial Intelligence, 2009.

[12] Ercan Canhasi and Igor Kononenko. Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization. Expert Systems with Applications, 41(2):535–543, 2014.

[13] Shufeng Xiong and Donghong Ji. Query-focused multi-document summarization using hypergraph-based ranking. Information Processing & Management, 52(4):670–681, 2016.

[14] Hadrien Van Lierde and Tommy WS Chow. Learning with fuzzy hypergraphs: A topical approach to query-oriented text summarization. Information Sciences, 496:212–224, 2019.

[15] Hadrien Van Lierde and Tommy WS Chow. Query-oriented text summarization based on hypergraph transversals. Information Processing & Management, 56(4):1317–1338, 2019.

[16] Aria Haghighi and Lucy Vanderwende. Exploring content models for multi-document summarization. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 362–370, 2009.

[17] Chao Shen, Tao Li, and Chris HQ Ding. Integrating clustering and multi-document summarization by bi-mixture probabilistic latent semantic analysis (plsa) with sentence bases. In Twenty-Fifth AAAI Conference on Artificial Intelligence, 2011.

[18] Yutong Wu, Yuefeng Li, and Yue Xu. Dual pattern-enhanced representations model for query-focused multi-document summarisation. Knowledge-Based Systems, 163:736–748, 2019.

[19] Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. Compressive document summarization via sparse optimization. In Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015.

[20] Guy Feigenblat, Haggai Roitman, Odellia Boni, and David Konopnicki. Unsupervised query-focused multi-document summarization using the cross entropy method. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 961–964, 2017.

[21] Haggai Roitman, Guy Feigenblat, Doron Cohen, Odellia Boni, and David Konopnicki. Unsupervised dual-cascade learning with pseudo-feedback distillation for query-focused extractive summarization. In WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020, pages 2577–2584, 2020.

[22] Reuven Y Rubinstein and Dirk P Kroese. The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning, volume 133. Springer, 2004.

[23] Xiaojun Wan and Jianmin Zhang. Ctsum: extracting more certain summaries for news articles. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, pages 787–796, 2014.

[24] Shulei Ma, Zhi-Hong Deng, and Yunlun Yang. An unsupervised multi-document summarization framework based on neural document model. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1514–1523, 2016.

[25] Sheng-hua Zhong, Yan Liu, Bin Li, and Jing Long. Query-oriented unsupervised multi-document summarization via deep learning model. Expert systems with applications, 42(21):8146–8155, 2015.

[26] Mahmood Yousefi-Azar and Len Hamey. Text summarization using unsupervised deep learning. Expert Systems with Applications, 68:93–105, 2017.

[27] Hal Daumé III and Daniel Marcu. Bayesian query-focused summarization. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pages 305–312, 2006.

[28] John M Conroy, Judith D Schlesinger, and Jade Goldstein Stewart. Classy query-based multi-document summarization. In Proceedings of the document understanding conference, 2005.

[29] Asli Celikyilmaz and Tur Dilek Hakkani. A hybrid hierarchical model for multi-document summarization. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 815–824, 2010.

[30] You Ouyang, Wenjie Li, Sujian Li, and Qin Lu. Applying regression models to query-focused multi-document summarization. Information Processing & Management, 47(2):227–237, 2011.

[31] Logan Lebanoff, Kaiqiang Song, and Fei Liu. Adapting the neural encoder-decoder framework from single to multi-document summariza-

tion. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4131–4141, 2018.

[32] Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1074–1084, 2019.

[33] Tetsuya Sakai et al. A comparative study of deep learning approaches for query-focused extractive multi-document summarization. In 2019 IEEE 2nd International Conference on Information and Computer Technologies (ICICT), pages 153–157. IEEE, 2019.

[34] Pengjie Ren, Zhumin Chen, Zhaochun Ren, Furu Wei, Liqiang Nie, Jun Ma, and Maarten De Rijke. Sentence relations for extractive summarization with deep neural networks. ACM Transactions on Information Systems (TOIS), 36:1–32, 2018.

[35] Pengjie Ren, Zhumin Chen, Zhaochun Ren, Furu Wei, Jun Ma, and Maarten de Rijke. Leveraging contextual sentence relations for extractive summarization using a neural attention model. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 95–104, 2017.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

[37] Yumo Xu and Mirella Lapata. Coarse-to-fine query focused multi-document summarization. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3632–3645, 2020.

[38] Ramakanth Pasunuru, Asli Celikyilmaz, Michel Galley, Chenyan Xiong, Yizhe Zhang, Mohit Bansal, and Jianfeng Gao. Data augmentation for abstractive query-focused multi-document summarization. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 13666–13674, 2021.

[39] Yumo Xu and Mirella Lapata. Document summarization with latent queries. Transactions of the Association for Computational Linguistics, 10:623–638.

[40] Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. Anaphora and coreference resolution: A review. Information Fusion, 59:139–162, 2020.

[41] Saliha Azzam, Kevin Humphreys, and Robert Gaizauskas. Using coreference chains for text summarization. In Coreference and Its Applications, 1999.

[42] Josef Steinberger, Massimo Poesio, Mijail A Kabadjov, and Karel Ježek. Two uses of anaphora resolution in summarization. Information Processing & Management, 43(6):1663–1680, 2007.

[43] Constantin Orăsan. The influence of pronominal anaphora resolution on term-based summarisation. In Recent Advances in Natural Language Processing V, pages 291–300. 2009.

[44] Christian Smith, Henrik Danielsson, and Arne Jönsson. A more cohesive summarizer. In Proceedings of COLING 2012: Posters, pages 1161–1170, 2012.

[45] Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. Towards coherent multi-document summarization. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2013.

[46] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, pages 280–290. Association for Computational Linguistics, 2016.

[47] Yuval Kirstain, Ori Ram, and Omer Levy. Coreference resolution without span representations. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 14–19. Association for Computational Linguistics, 2021.

[48] Shon Otmazgin, Arie Cattan, and Yoav Goldberg. LingMess: Linguistically informed multi expert scorers for coreference resolution. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2752–2760, Dubrovnik, Croatia, 2023. Association for Computational Linguistics.

[49] Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task, pages 1–27, June 2011.

[50] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In Joint Conference on EMNLP and CoNLL - Shared Task, pages 1–40, 2012.

[51] Saad Ezzini, Sallam Abualhaija, Chetan Arora, and Mehrdad Sabetzadeh. Automated handling of anaphoric ambiguity in requirements: a multi-solution study. In Proceedings of the 44th International Conference on Software Engineering, pages 187–199, 2022.

[52] Regina Barzilay and Noemie Elhadad. Inferring strategies for sentence ordering in multidocument news summarization. Journal of Artificial Intelligence Research, 17:35–55, 2002.

[53] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference, 2015.

[54] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, 2018.

[55] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 1–14, 2017.

[56] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation, 10(7):1895–1923, 1998.

[57] Daniel Deutsch, Rotem Dror, and Dan Roth. A statistical analysis of summarization evaluation metrics using resampling methods. Transactions of the Association for Computational Linguistics, 9:1132–1146, 2021.

[58] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, 2014.

[59] Salima Lamsiyah, Abdelkader El Mahdaouy, Bernard Espinasse, and Said Ouatik El Alaoui. An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings. Expert Systems with Applications, page 114152, 2020.

[60] Stephen E Robertson, Steve Walker, Susan Jones, et al. Okapi at trec-3. 1995.

[61] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683, 2019.

SALIMA LAMSIYAH graduated from the Faculty of Science Dhar El Mahraz at Sidi Mohamed Ben Abdellah University with a degree in Computer Science and Mathematics in 2014. In 2016, she obtained her Master's Degree in Information Systems, Networks, and Multimedia from the same faculty. She then completed her Ph.D. in Computer Science from Ibn Tofail University, Morocco in collaboration with Aix Marseille University in December 2021.

Currently, she is a Postdoctoral Researcher in the field of Natural Language Processing and Machine Learning at the University of Luxembourg. Her research interests include Machine Learning, Natural Language Processing, and Deep Learning.

ABDELKADER EL MAHDAOUY graduated from the Faculty of Science Dhar El Mahraz of Sidi Mohamed Ben Abdellah University in Computer Science, in 2012. Then, he received his Ph.D. in Computer Science in a joint program between Grenoble Alps University and Sidi Mohamed Ben Abdellah University in December 2017.

In September 2020, he joined Mohammed VI Polytechnic University (UM6P) as a Postdoctoral Researcher in Artificial intelligence (AI) for Cybersecurity. Currently, he is a Scientist at UM6P, Ben Guerir, Morocco. His research interests include Machine Learning, Data Science, and Natural Language Processing.

CHRISTOPH SCHOMMER studied Artificial Intelligence at the University Saarbrücken before working at IBM for 8 years. During the same period, he completed a Ph.D. in Medical Informatics at the Goethe University Frankfurt/Main.

In October 2003, Christoph was appointed Associate Professor at the University of Luxembourg. Today, Christoph heads the MINE research group and the ACC laboratory. His interests lie in Artificial Intelligence and at the intersection of Machine Learning and Data Science. He is an internationally recognized scientific reviewer for the Leibniz Association, Springer, IEEE and served as PC member at more than 100 international conferences (such as IJCAI, AAMAS, CogSci, ECML, and others). He regularly organizes lecture series/Ph.D. workshops and is the author of approx. 100 scientific papers. He supervised 30 Ph.D. projects in Luxembourg, Turin and London and currently oversees 12 Ph.D. projects. He has taught a total of 170 courses at the University of Luxembourg (Computer Sciences, Mathematics, Finance) and at several other universities in the EU, Beijing (Tsinghua) and Singapore (SUTD). Christoph maintains contacts with industry and the National Ethics Council and is a member of the ACM, the Cognitive Science Society, and the Deutsche Gesellschaft für Kognitionswissenschaft.

TABLE 7. Example of the generated summary for **Cluster D654f** from DUC'2005 dataset using our method CohQFMDS-SBERT-Sum based on Scenario 2 where the broken pronouns are replaced by their corresponding entities.

Query

- What are the advantages and disadvantages of same-sex schools?

Generated summary

- $\mathbf{S}_1$ The academic debate over the quality of single-sex education as compared with a co-educational setting, however, has leaned towards the conclusion that, at least for girls, single-sex education is better.

- $\mathbf{S}_2$ **This year's survey (It)** suggests that, with impressive exceptions, day schools do better than traditional boarding schools and that single-sex schools do better than co-educational schools.

- $\mathbf{S}_3$ **Single-sex schools (They)** achieved better results than fully mixed schools in both the state and independent sectors.

- $\mathbf{S}_4$ Single-sex education was viewed as a barrier to the successful socialization of adolescents, said Valerie Lee, a professor at the University of Michigan in Ann Arbor.

- $\mathbf{S}_5$ Lee's research shows that in schools where students are of the same socioeconomic status, "single-sex schools seem to be advantageous to the women that attend them".

- $\mathbf{S}_6$ For both boys and girls, **Valerie Lee (she)** said, "all of the statistically significant single-sex school effects were positive" and there were "few negative effects."

- $\mathbf{S}_7$ **Mr John Sugden (He)** backs mixed schools on educational grounds as well: 'If you put the two sexes together you are bound, if you get the balance right, to get a better education. Otherwise, you are cutting out half the human race.'

- $\mathbf{S}_8$ Parents of potential future students have said they would like a co-ed school in the area, Clarke said, and offering the choice between co-ed and single-sex education was a motivating factor in the decision.

- $\mathbf{S}_9$ The FT scores for single-sex schools are identical for boys and girls (1.09 in independents and 1.04 in grammar schools), whereas mixed independent schools averaged 1.01, with grammar schools on 0.95.

Gold summary

- Research shows that in schools where students are of the same socio-economic status, single-sex schools seem to be advantageous to girls and that for boys it does not seem to make much of a difference whether they go to single-sex or co-ed schools.

- Advocates of single-sex schools cite better test score.

- Surveys in English schools indicated that the best schools, academically, were single-sex schools.

- In 1992, 25 of the top 30 state schools were single-sex schools; and in 1993, 29 of the top 30 were single-sex schools.

- Advocates claim that girls in single-sex schools have more confidence and this contributes to their high academic performance.

- It is claimed that girls in single sex-schools can support each other.

- In addition, girls do not have to conform to gender stereotypes that they should not study science or mathematics.

- Another argument for single-sex schools is that girls are not distracted by boys and boys are not distracted by girls.

- Research shows that in co-ed colleges, men dominate classroom discussions. There is much debate about the significance of single-sex schools getting better examination grades.

- Research suggests that at the age of 16 girls are more mature and intellectually more advanced than boys, who start to mature at 17 or 18.

- Girls generally outperform girls on the GCSE – the main examination for 16-year-olds in England and Wales.

- Boys, however, catch up when A-level examinations, used for college admission, are taken.

- Opponents of single-sex education viewed it as a barrier to successful socialization of adolescents.