Received: 13 September 2022 | Revised: 27 April 2023 | Accepted: 15 September 2023

DOI: 10.1111/exsy.13462

ORIGINAL ARTICLE

Expert Systems WILEY

SDbQfSum: Query-focused summarization framework based on diversity and text semantic analysis

Muhidin Mohamed[1] | Mourad Oussalah[2] | Victor Chang[1]

[1]Department of Operations and Information Management, Aston University, Birmingham, UK

[2]Faculty of ITEE, CMVS, University of Oulu, Oulu, Finland

Correspondence
Muhidin Mohamed and Victor Chang, Department of Operations and Information Management, ABS, Aston University, Birmingham, B4 7ET, UK.
Email: m.mohamed10@aston.ac.uk and v.chang1@aston.ac.uk

Abstract

Query-focused multi-document summarization (Qf-MDS) is a sub-task of automatic text summarization that aims to extract a substitute summary from a document cluster of the same topic and based on a user query. Unlike other summarization tasks, Qf-MDS has specific research challenges including the differences and similarities across related document sets, the high degree of redundancy inherent in the summaries created from multiple related sources, relevance to the given query, topic diversity in the produced summary and the small source-to-summary compression ratio. In this work, we propose a semantic diversity feature based query-focused extractive summarizer (SDbQfSum) built on powerful text semantic representation techniques underpinned with Wikipedia commonsense knowledge in order to address the query-relevance, centrality, redundancy and diversity challenges. Specifically, a semantically parsed document text is combined with knowledge-based vectorial representation to extract effective sentence importance and query-relevance features. The proposed monolingual summarizer is evaluated on a standard English dataset for automatic query-focused summarization tasks, that is, the DUC2006 dataset. The obtained results show that our summarizer outperforms most state-of-the-art related approaches on one or more ROUGE measures achieving 0.418, 0.092 and 0.152 in ROUGE-1, ROUGE-2, and ROUGE-SU4 respectively. It also attains competitive performance with the slightly outperforming system(s), for example, the difference between our system's result and best system in ROUGE-1 is just 0.006. We also found through the conducted experiments that our proposed custom cluster merging algorithm significantly reduces information redundancy while maintaining topic diversity across documents.

KEYWORDS

query-focused summarization, query-relevance, semantic role labeling, sentence centrality, sentence similarity

1 | INTRODUCTION

Nowadays, unstructured textual data is the dominant component of big data, as can be noticed from electronic news outlets, social media platforms, online customer product/service reviews, exchanged emails, electronic magazines, journals, books, and so forth. Although, all these form a

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2023 The Authors. *Expert Systems* published by John Wiley & Sons Ltd.

rich information and raw data source, it will be unlikely for the readers to effectively go through and make use of all this textual information without some form of automated management systems. This is why text summarization technologies are more widely needed than ever before. Summarization is a sub-task of natural language processing (NLP), a branch of AI that studies the development of computer systems for understanding human languages. It aims to reduce source text documents to brief substitute summaries that either answer a user question or retain the gist of the document.

Text summarization can be briefly classified on various bases. To start with, we can differentiate between *extractive* and *abstractive* summarization approaches based on the purpose of the summary to be produced. Extraction-based approaches are the dominant and most well-established ones, although, abstraction-based methods have recently started to achieve some success with the growth of neural based summarization (Gupta & Gupta, 2019). In extractive summarization, document contents, for example, sentences, are scored and ranked on the basis of some predefined salient indicators. Top ranked sentences, up to a given threshold, are then extracted as the summary. Depending on the source input text, we distinguish two summarization tasks: *single document summarization (SDS)*, and *multi-document summarization (MDS)*. In addition, if the purpose of generating a summary from an input document is to gain the gist or the overall meaning of that document, then the respective method is called *generic* or *topic-focused* summarization. If the summary is to answer a user information need or a given question (query), it is called a *query-focused* summarization. This study falls in the category of extractive query-focused multi-document summarization as the developed approach is meant to answer complex user queries in the form of summaries extracted from related document sets.

Research on text summarization has achieved good progress in methods, applications, and languages since Luhn's pioneering work in 1958 (Luhn, 1958), but significant number of challenges still exist in the field to produce meaningful, coherent, easily readable summaries (El-Kassas et al., 2021). And unlike SDS (based on single source document), MDS (based on several related documents) has a number of additional challenges associated with it. This includes the differences and similarities across related document clusters, the high degree of redundancy inherent in extracts created from multiple sources, and the small source-to-summary compression ratio. This study attempts to address some of these challenges by investigating the following key research questions.

1. What are the current challenges and limitations in extractive multi-document summarization, and how can we address them? This is discussed in sections 3 & 4 of this paper.
2. Can the topic diversity of a multi-document summary be improved by reducing the information overlap at the pre-processing and summary extraction stages? This is addressed in Section 4 and the experiment part of this paper.
3. Will the use of text semantic representation methods produce reliable features for scoring document sentences and establishing query relevance? This is addressed in Section 6 of this manuscript.

Especially, we propose a query-focused summarization framework built on effective text semantic representation techniques and a diversity principle in this paper. The framework uses the entire Wikipedia database as an external knowledge. Specifically, we apply a diversity-based approach to retrieve diverse concepts in the query reply summary, and semantic analysis techniques to enrich and improve sentence (and query) scoring features. The work is partially motivated by the improvements in measuring short text semantic similarities in our previous Wikipedia graph based single and multi-document text summarization approach (Mohamed, 2016; Mohamed & Oussalah, 2019). Besides, relevant literature confirms that using language semantics has the potential for building improved summarization approaches, which provides motivation for further research investigation in this area (Nenkova & McKeown, 2012). This study aims to contribute to uncovering some of the query-focused summarization challenges, such as information redundancy, query-relevance, and topic coverage using enriched sentence importance, feature diversity, and Wikipedia lexical database. More specifically, the contributions of this paper are threefold:

1. First, we adopt a two-stage diversity technique, a custom cluster merging method and maximal marginal relevance (MMR) algorithm, to address information redundancy, query-relevance, and topic centrality challenges inherent in multi-document summarization tasks.
2. Second, we use semantic representation techniques to extract effective sentence features for modeling query-relevance, sentence centrality, and diversity factors.
3. Third, we employ extracted semantic features to implement a semantic diversity feature based query-focused summarizer (SDbQfSum), which we evaluated on DUC2006 dataset.

The rest of the paper is organized as follows. Sections 2 and 3 present research objectives and related works. Sections 4 and 5 briefly introduce the applied diversity and semantic analysis and representation techniques. Detailed description of the proposed summarization approach, and its experimental evaluation are discussed in sections 6 and 7, respectively. Finally, a conclusion is given in Section 8.

2 | RESEARCH OBJECTIVES

While text summarization is a difficult task in general, generating a summary that effectively answers a user question from multiple related documents is even more complex in several ways. This includes (i) making the extracted summary relevant to the user question both lexically and

14680394, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13462 by Test, Wiley Online Library on [02/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

semantically (query-relevance); (ii) ensuring that all opinions expressed in the different documents are represented in the produced summary (coverage); (iii) Eliminating (or minimizing) information repetitions from the various source documents in the length restricted synopsis (diversity).

In this work, we use proven and effective semantic analysis and representation methods to address proceeding challenges and design a query-focused summarizer. Extracting the underlying meaning of the text is intended to create useful sentence semantic features for capturing the aforementioned relevance, coverage and diversity aspects. The overall proposed summarization approach is technically described throughout this paper. The consideration of these factors underpinned with text semantic features and Wikipedia knowledge have improved the query-focused summarization, as shown in the presented experimental evaluation. In a nutshell, the specific objectives of this paper are as follows:

1. To extend our SRL-ESA Wikipedia based approach (Mohamed & Oussalah, 2019) to query-focused redundancy aware summarization built on text semantic features and summary diversity principles.
2. To develop a two-staged diversity technique that includes a document cluster merging approach in the preprocessing stage and MMR-like method in the sentence ranking and summary extraction stages.
3. To validate the proposed query-focused summarization on a benchmark dataset (DUC2006) with a comparison to state-of-the-art methods.

3 | RELATED WORKS

Overall, any extractive summarization task undergoes three essential steps: intermediate representation of the source text, document content scoring (e.g., sentences), and selection of the highest scored contents as a summary (Nenkova & McKeown, 2012). Some of the most widely used document representation techniques for identifying salient content in text summarization task include semantic approaches (Bidoki et al., 2020; Mohamed & Oussalah, 2019), knowledge-based methods (Abdi et al., 2017; Mohamed & Oussalah, 2015), graph-based techniques (Canhasi & Kononenko, 2014; Mohamed & Oussalah, 2019), feature-driven approaches (Mutlu et al., 2019; Ouyang et al., 2011), and diversity-based methods (Binwahlan et al., 2010; Gambhir & Gupta, 2017; Luo et al., 2013). With some differences in wording and terminology, query-focused text summarization systems have been widely framed in the literature using the concepts of relevance, coverage and novelty with diversity awareness (Binwahlan et al., 2010; Carbonell & Goldstein, 1998; Luo et al., 2013; Mohamed & Oussalah, 2015). These factors (novelty, diversity & coverage) are also used in several other related research areas, for instance, ranking recommended items of recommendation systems (Zangerle & Bauer, 2022; Zehlike et al., 2022), retrieving documents in an information retrieval system (Meng & Shen, 2018), and so forth.

Carbonell and Goldstein (1998) proposed the pioneering *MMR* algorithm for maximising diversity and minimising redundancy in automatically generated summaries. In other words, their algorithm simultaneously ensures that the produced summary sentences are highly query-relevant, cover distinct concepts in the source document, and minimise redundancy. Relevant literature confirms the effectiveness of the MMR algorithm for query-based summarization tasks (Murray et al., 2005; Verberne et al., 2020). The MMR algorithm and its derivatives have been widely adopted and/or used in text summarization tasks for the same purpose (Binwahlan et al., 2010; Mohamed & Oussalah, 2015; Verberne et al., 2020). For example, Verberne et al. (2020) proposed a comparison of two automatic summarization methods for discussion threads: a supervised query-independent approach built on post features, and an unsupervised MMR-based method constructed using query-dependent features. Although, *MMR* has been previously acknowledged to be a successful method for query-focused conversation summarization, authors of this particularly study found that the method based on post features outperforms the MMR-based method (Murray et al., 2005; Verberne et al., 2020). This has been attributed to the fact that the *MMR* algorithm is an unsupervised approach compared to the supervised post-feature approach. In addition, Binwahlan et al. (2010) introduced a modified version of the *MMR*algorithm called Maximum Marginal Importance (MMI) for generic summarization. Their diversity-based summarization method is built on a range of sentence features including sentence centrality and similarity with title/first sentence. Other less common methods of managing redundancy in multi-document summarization include sentence similarity measures (Alguliev et al., 2013; Mosa et al., 2019). Besides, modeling sentence importance based on semantic relevance was also found to improve summarization (Wang et al., 2016)

The use of different semantic representation methods for extractive text summarization has been growing recently (Abdi et al., 2017; Bedi et al., 2022; Bidoki et al., 2020; El-Kassas et al., 2021). These semantic summarization methods consist of several loosely classified categories. They include approaches based on semantics derived from linguistic knowledge, such as the work of Abdi et al. (2017) who proposed a query-focused summarization model using a combination of word semantic relations and syntactic composition to identify key document sentences. In particular, their system uses a graph representation of documents whose nodes are linked by the sentence semantic similarities with WordNet-assisted word expansion. Besides, Bedi et al. (2022) proposed an unsupervised extractive summarizer based on semantic similarity and key-phrase extraction, which was applied to biomedical datatset for validation. Distributional approaches such as explicit semantic analysis (ESA) and latent semantic analysis (LSA) have been successfully employed in text summarization tasks as well (Al-Sabahi et al., 2018; Nenkova & McKeown, 2012; Ozsoy et al., 2011). Examples of summarization proposals built on distributional semantics include the works of Ozsoy et al. (2011) and Al-Sabahi et al. (2018) both of whom suggested LSA based summarization approaches for Turkish and Arabic languages respectively. ESA is a less common approach but has been pioneered by a few previous works, including that of Sankarasubramaniam et al. (2014) which uses the algorithm to

14680394, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13462 by Test, Wiley Online Library on [02/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

represent the meaning of words. The approach has also been adopted in our previous work for generic single and multi-document summarization (Mohamed & Oussalah, 2019). Another class of semantic summarizers refer to methods built on word semantic vectors, which underpins the recent boom in neural based summarization (Bidoki et al., 2020; Dong, 2018).

In this study, we use the strengths of text semantic representations to model a diversity-aware query-focused summarizer using a set of relevance, centrality and novelty features. The semantic representation techniques used in this paper (i.e., semantic role labelling and explicit semantic analysis) have been used together for the first time in our previous work (Mohamed & Oussalah, 2019). While the SRL-ESA combination is re-used in this study, there are some important differences between the preceding and current research works. Especially, the semantic representations were used in (Mohamed & Oussalah, 2019) to improve document similarity for generic graph-based single and multi-document summarization tasks via enhancing the measurement and computation of sentence semantic similarities. However, the semantic techniques were applied to a different summarization task (i.e., query-focused) in this work with the objective of extracting improved text semantic features. In other words, this study builds on the previous work in adopting the SRL-ESA Wikipedia-based method to realize a query-focused redundancy aware summarization model built on text semantic features and summary diversity principle.

4 | HANDLING THE DIVERSITY CHALLENGE

The term diversity, in the context of text summarization, refers to the reduction of redundancy and inclusion of the largest number of distinct concepts in the produced summary. In other words, it attempts to select non-redundant dissimilar summary sentences, based on combined feature functions of query-relevance and summary diversity. Information novelty and redundancy avoidance are two primary challenges in multi-document summarization, as a single summary is sought from different documents that describe the same topic. Especially, the process of summarising a cluster of articles creates such challenges, primarily due to the effect of merging different descriptions of the same concepts usually authored by different writers on various news sources. To tackle these challenges, several approaches have been suggested. This includes the use of sentence similarity (Alguliev et al., 2013; Mosa et al., 2019), *MMR*-like reasoning and its derivatives (Binwahlan et al., 2010; Carbonell & Goldstein, 1998; Gambhir & Gupta, 2017), or clustering algorithm-based approaches (Alguliyev et al., 2019; Ferreira et al., 2014). In this work, we employ a two-staged diversity technique, a novel document cluster merging approach in the preprocessing stage, and *MMR*-like reasoning algorithm in the summary ranking and extraction stage (Carbonell & Goldstein, 1998; Mohamed & Oussalah, 2019).

4.1 | Merging document clusters

This technique was first designed and used in our previous work on generic graph-based text summarization (Mohamed & Oussalah, 2019) for mitigating the redundancy challenge. It is a simple algorithm that works as follows. Firstly, source document clusters to be summarised are merged together to form a single flattened document for each cluster, while arranging its sentence corpus in the order of the source documents' timeline. Secondly, an iterative algorithm for removing similar sentences is applied to exclude repeated information. This is achieved by computing the similarity of each sentence with the other cluster sentences and removing those which are highly similar to it. The result is a unified cluster document with minimized content repetition.

Formally speaking, if $\mathbb{D} = \{D_1, D_2, D_3, \ldots, D_M\}$ is a set of M documents on the same topic to be summarised, we join all sentences of the document set getting a flattened cluster, $C = \{s_1, s_2, s_3, s_4, s_4, s_4, \ldots, s_n\}$, where n is the total number of cluster sentences. Next, cluster sentences, s_i, i=1 to n, are filtered in order to reduce each cluster of very similar sentences to a single sentence. For illustrative purpose, assume that sentence one (s_1) in Figure 1 is compared to the rest of the cluster sentences and is found to be highly similar to sentences 6, 8 & 12 (thick lines). To reduce the redundancy, the algorithm drops these highly overlapping sentences in their meaning and reduces the cluster to the single source sentence against which the rest were compared (sentence 1 in this case). In other words, the filtering process iterates over the cluster sentences and forwardly keeps the current one to be included in the resulting unified non-redundant k cluster sentences with the result of eliminating ($n - k$, $n \geq k$) sentences from the corpus at the end.

4.2 | Maximum marginal relevance

Maximum Marginal Relevance or *MMR* is an algorithm designed to minimize information redundancy while maximizing diversity in text summarization and information retrieval (Carbonell & Goldstein, 1998). It enables the inclusion of diverse concepts and opinions from the source document(s) in the extracted summary by maximizing summary diversity and query relevance at the same time. Maximum marginal relevance in the query-focused summarization context means a sentence is very relevant to the user-query and does not repeat information already included

14680394, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13462 by Test, Wiley Online Library on [02/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

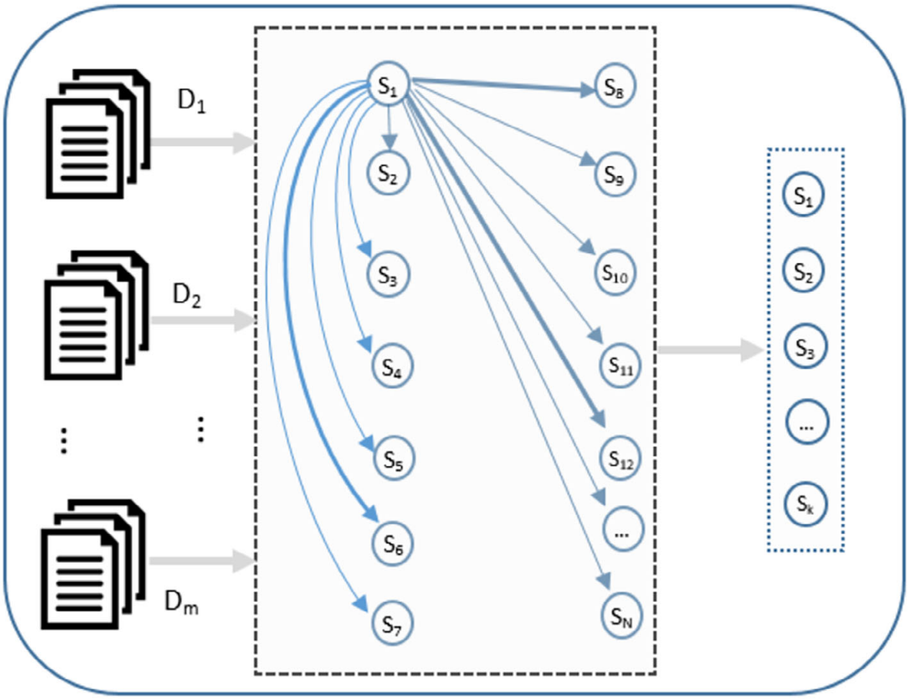


FIGURE 1 Merging cluster documents with redundancy removal.

in the summary. Intuitively, devising an anti-redundancy mechanism with restricted summary length (as required for the DUC2006 and other datasets) can be perceived to entail diversity in text summarization. Equation (1) provides a formal mathematical definition of the *MMR* algorithm.

$$MMR(SC,Q,R,Sum) = \underset{S_i \in R|Sum}{\arg} \max\left[\lambda Sim_1(S_i,Q) - (1-\lambda)\max_{S_j \in Sum} Sim_2(S_i,S_j)\right] \quad (1)$$

The first and the second functions of the right hand side in (1) are for the query-relevance and dissimilarity maximisation respectively, *SC* represents cluster sentences; *Q* denotes the query; *R* is the list of scored and ranked sentences, *Sum* indicates the sentences already selected from *R* for summary; *R*|*Sum* is the remaining unselected sentences in R, while Sim_1 and Sim_2 are the applied similarity functions. The λ parameter in the equation is a factor used to weight the contribution of each of the two components in the combined *MMR* similarity formula. With $\lambda = 1$, the measure calculates the strength of the maximum query-relevance, and the maximal sentence diversity with $\lambda = 0$. Besides, one can use λ values between 0 and 1 to strike the right balance between relevance and diversity.

5 | APPLIED SEMANTIC ANALYSIS METHODS

5.1 | Semantic role labeling

Semantic role labeling (aka SRL) is a method for parsing short texts such as sentences. It recognizes the semantic role that each word in a sentence occupies with respect to its predicate verb(s). The semantic roles (aka frame elements) form the building blocks of a semantic frame, and word meanings are defined in relation to semantic frames. Several well-established manually annotated lexical databases, including FrameNet (Baker et al., 1998) and Probank (Fillmore & Baker, 2012), are specifically created for linking semantic frames to word meanings. Investigation into the development of automatic semantic role labeling systems has been a recently growing NLP researched topic (He et al., 2017). One of the pioneering works in automatic sentence semantic parsing was proposed by Gildea and Jurafsky (2002). Their algorithm was built on a statistical classifier trained on annotated data from FrameNet. Likewise, Collobert et al. (2011) recently proposed a multi-task deep learning based system designed for different NLP predictions such as part-of-speech tagging, chunking, named entity recognition, and semantic role labelling. Their prediction models were designed such that they learn from large, predominantly un-annotated, training data. The outcome of this work was open-source software, called SENNA, which we used for predicting frame elements in our work. SENNA's unique strengths include high speed performance and minimal computational requirements. In essence, SRL identifies all argument terms that fill semantic roles for sentence predicate

14680394, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13462 by Test, Wiley Online Library on [02/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

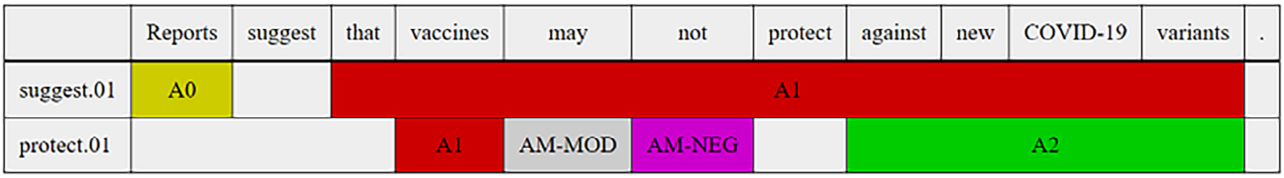
verb(s) and assigns their corresponding semantic role tags. The semantic parsing with SRL can be useful in extracting information on *who did what to whom*, *when*, *where* and *why*, as demonstrated in Example 1.

Example 1. Reports suggest that vaccines may not protect against new COVID-19 variants.

Figure 2 demonstrates the sentence in Example 1 parsed with the Lund Semantic Role Labeler*. The SRL parser identifies the sentence predicate verbs and associated arguments. There are two groups of SRL arguments: Core semantic arguments, for example, Agent (subject), Theme (direct object), Instrument, and so on; and non-core (aka adjunctive) arguments, for example, Purpose, Location, Temporal, Negation, Manner, Extent, Cause, and so forth. Table 1 shows the list of semantic role arguments and their associated tags. The figure shows that the sentence has two predicate verbs: *suggest* and *protect*, and hence, two semantic frames. As in Table 2, the *suggest* predicate verb has two semantic role arguments: subject (*reports*) and object (*that vaccines may not protect against new COVID-19 variants*). On the other hand, the semantic role-set of the *protect* predicate includes object (*vaccines*), indirect object (*against new COVID-19 variants*), and two adjunctive arguments: model verb (*may*) and negation (*not*).

5.2 | Explicit semantic analysis

Explicit Semantic Analysis (aka ESA) is a knowledge-based vector representation of text for computing semantic relatedness proposed by Gabrilovich and Markovitch (Gabrilovich & Markovitch, 2009). The algorithm translates words in a text to concept vectors extracted from Wikipedia on the assumption that the encyclopedia articles represent linguistic concepts, that is, mapping of text terms to their corresponding concepts is in this context perceived as meaning representation. Technically, the ESA algorithm is used to represent input texts by building an inverted index of weighted Wikipedia concepts using its entire corpus. This is achieved by iterating over each term in a text and interpreting it to its accommodating natural concepts. Precisely, the semantic relatedness of any two short texts is calculated by comparing their translated concept vectors using cosine similarity. Figure 3 provides a high level illustration of the explicit semantic analysis method. That is, the semantic relatedness of two sentences, say, *Sentence*$_1$ and *Sentence*$_2$ boils down to the following. First, it calls the semantic interpreter over each token of either *Sentence*$_1$ or *Sentence*$_2$ iteratively. Second, it retrieves its corresponding entry from the inverted index. Third, using the Wikipedia lexical database,



	Reports	suggest	that	vaccines	may	not	protect	against	new	COVID-19	variants	.
suggest.01	A0		A1									
protect.01				A1	AM-MOD	AM-NEG		A2				

FIGURE 2 Example 1 semantically parsed with SRL.

TABLE 1 Semantic role arguments and associated labels.

Core arguments		Non-core arguments	
Label	**Modifier**	**Label**	**Modifier**
V	Verb	AM-DIR	Direction
A0	Subject	AM-ADV	Adverb
A1	Object	AM-LOC	Location
A2	Indirect object	AM-TMP	Temporal marker
A3	Start point	AM-MNR	Manner
A4	End point	AM-DIS	Discourse marker
A5	Direction	AM-PRP	Purpose
—	—	AM-NEG	Negation
—	—	AM-EXT	Extent
—	—	AM-PNC	Proper noun
—	—	AM-MOD	Model verb

14680394, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13462 by Test, Wiley Online Library on [02/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 Verb-argument pairs for the case of Figure 2.

Arguments/Verbs	Protect	Suggest
A0	Reports	—
A1	That vaccines…	Vaccines
A2	—	Against new COVID-19 variants
AM-MOD	—	May
AM-NEG	Not	—

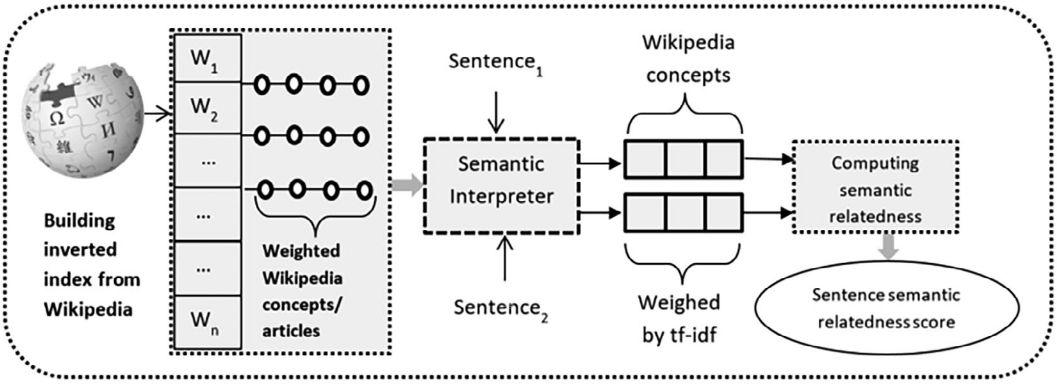


FIGURE 3 Computing sentence semantic relatedness with ESA algorithm.

we represent the term semantics by the corresponding TF-IDF (term frequency-inverse document frequency) weighted concept vector inferred from Wikipedia.

More formally, let $T = w_i (i = 1 \; to \; m)$ be the input text and let $\overrightarrow{K_{w_i}}$ be the inverted index entry for token w_i, where K_{w_i} quantifies the strength of the association between w_i and the Wikipedia concept set $CS = \{c_1, c_2, ..., c_N\}$. Therefore, the semantic interpretation for T is represented by the vector $V = [v_1, v_2, ..., v_N]$, where $v_j (j = 1 \; to \; N)$ evaluates the connection between concept c_j and the text T, and is defined as $\sum_{w_i \in T} tf.idf_{w_i} * K_{w_i}$. Using the TF-IDF, each word of T can be assigned the weight:

$$tf.id\,f(w,a) = tf_{w,a} \cdot log \frac{N}{n_w} \tag{2}$$

The $tf_{w,d}$ in Equation (2) stands for the term frequency of word w in Wikipedia article a, n_w denotes the number of articles containing w, and N is the number of Wikipedia articles (English). The ESA algorithm computes the semantic relatedness from Wikipedia concept vectors representing the words of text T. Therefore, the semantic relatedness $SemRel(S_1, S_2)$ of the sentences S_1 and S_2 is computed by performing the cosine similarity on their respective semantic interpretation vectors V_1 and V_2 as in Equation (3).

$$SemRel(S_1, S_2) = \frac{V_1.V_2}{\|V_1\| \|V_2\|} \tag{3}$$

6 | PROPOSED SEMANTIC-BASED QUERY-FOCUSED TEXT SUMMARIZATION MODEL

In the context of this paper, the problem of query-focused multi-document summarization can be defined as follows. Given a set of related documents (C_i) each merged to form N sentences, $C_i = \{s_1, s_2, s_3, ..., s_N\}$, the aim is to obtain a subset summary Sum ($Sum \subseteq C_i$) which is query relevant (QR), optimises the scoring function F_i, and has the maximum cluster coverage (mCC_i); $S = \underset{i}{argmax}\{F_i | \; s_i \; is \; query \; relevant \; with \; mCC_i\}$. The detailed formulation of the query relevance (mCC_i) will be provided in the next subsections. Figure 4 shows an illustration of the proposed query-focused summarization model involving three main components, which are briefly described below.

1. **Pre-processing and semantic parsing**: This element of the system (top left of Figure 4) turns the raw textual data into semantic representations. It takes a set of related documents and user queries as inputs and produces their semantic representations. It includes diversification

14680394, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13462 by Test, Wiley Online Library on [02/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

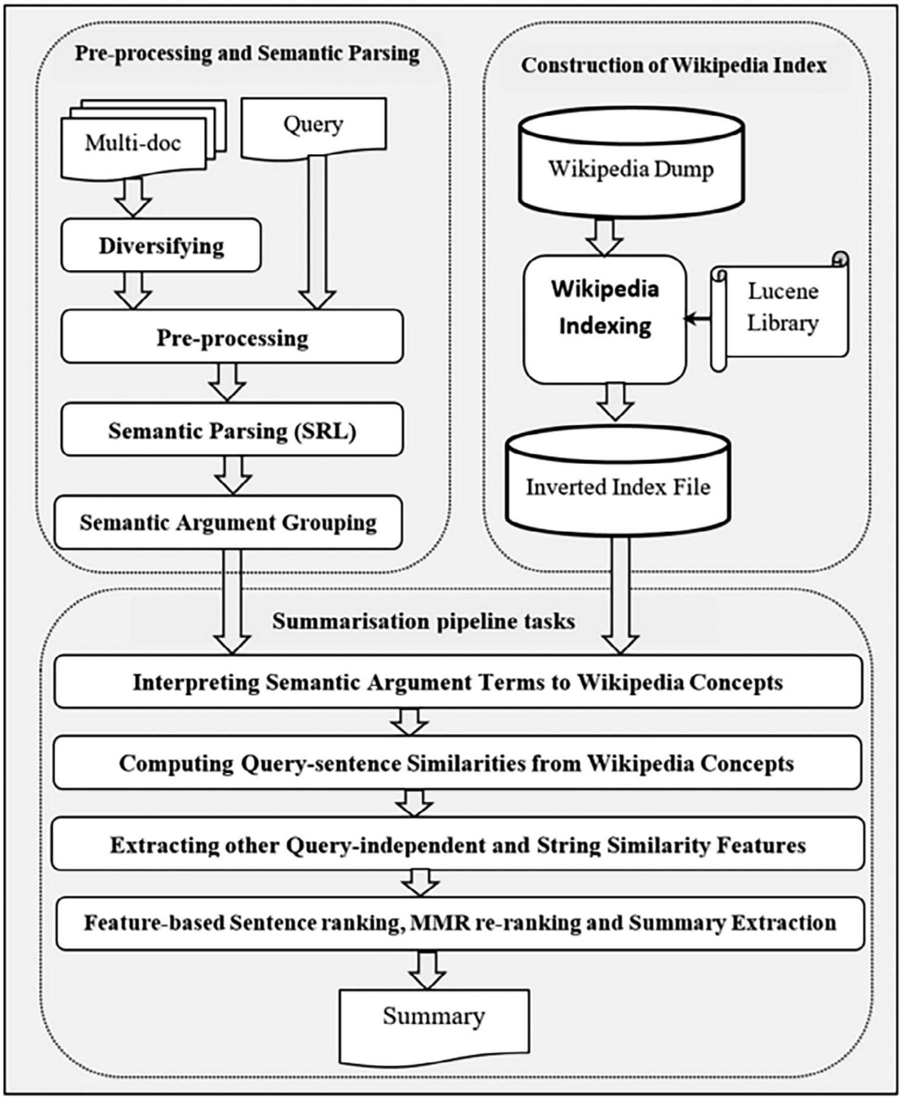

Pre-processing and Semantic Parsing
Multi-doc
Query
Diversifying
Pre-processing
Semantic Parsing (SRL)
Semantic Argument Grouping
Construction of Wikipedia Index
Wikipedia Dump
Wikipedia Indexing
Lucene Library
Inverted Index File
Summarisation pipeline tasks
Interpreting Semantic Argument Terms to Wikipedia Concepts
Computing Query-sentence Similarities from Wikipedia Concepts
Extracting other Query-independent and String Similarity Features
Feature-based Sentence ranking, MMR re-ranking and Summary Extraction
Summary

FIGURE 4 Proposed semantic diversity feature based query-focused summarization system (SDbQfSum).

(merging with redundancy removal, see Section 4.1) of each document cluster associated with a single topic to form one cluster document. Then, pre-processing is applied to the merged multi-documents and queries, followed by SRL-based semantic argument grouping. The pre-processing step involves applying standard NLP tasks such as document segmentation, sentence tokenization, part-of-speech tagging, word stemming and removing noise words. Besides, the purpose of the semantic parsing is to identify the semantic frames and their arguments while the semantic argument grouping is intended to collect all argument terms of the same semantic role in a sentence and link them to their modifier (cf. Figure 5).

2. **Construction of wikipedia index**: As alluded to, the summarization framework proposed in this paper relies on Wikipedia database as the external knowledge for the conceptual expansion of the text present in the experimental data. To this end, the entire English Wikipedia corpus is used to construct an inverted index file (top right of Figure 4) built with the adaptation of Apache Lucene Library[†]. The resulting inverted index file is what provides the mapping of the argument terms to the corresponding Wikipedia concepts using the ESA algorithm (Gabrilovich & Markovitch, 2009) as explained in Section 5.2.
3. **Summarization pipeline tasks**: After the preparation of semantic role-argument pairs along with the creation of a look up database for Wikipedia concepts, the implementation proceeds to the next phase. This stage comprises of core summarization tasks (bottom of Figure 4), which we discuss in the following sections. This multi-staged realization of the summarization system achieves a number of goals simultaneously. First, it enables efficient semantic representation of sentence texts with grouped semantic role-arguments for feature extraction. Second, it links document concepts to external common sense knowledge in Wikipedia for capturing meaning beyond the text, while achieving diversity in the produced summary at the same time. This step includes the application of the MMR algorithm for summary extraction, re-ranking and redundancy mitigation.

14680394, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13462 by Test, Wiley Online Library on [02/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

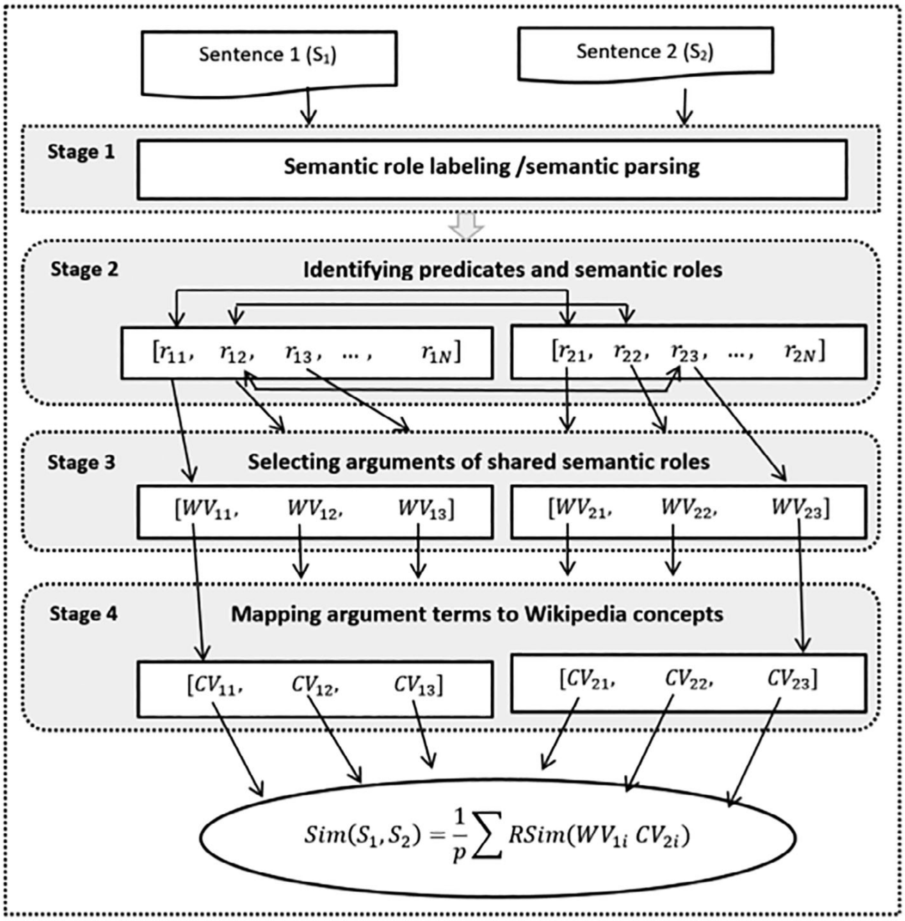

Sentence 1 (S_1)
Sentence 2 (S_2)
Stage 1
Semantic role labeling /semantic parsing
Stage 2
Identifying predicates and semantic roles
$[r_{11}, r_{12}, r_{13}, \ldots, r_{1N}]$
$[r_{21}, r_{22}, r_{23}, \ldots, r_{2N}]$
Stage 3
Selecting arguments of shared semantic roles
$[WV_{11}, WV_{12}, WV_{13}]$
$[WV_{21}, WV_{22}, WV_{23}]$
Stage 4
Mapping argument terms to Wikipedia concepts
$[CV_{11}, CV_{12}, CV_{13}]$
$[CV_{21}, CV_{22}, CV_{23}]$
$Sim(S_1, S_2) = \frac{1}{p} \sum RSim(WV_{1i}\ CV_{2i})$

FIGURE 5 Sentence semantic similarity using SRL-ESA based approach.

6.1 | The semantic similarity measure

A core component of the proposed summarization approach is the computation of the semantic similarity for query-relevance, sentence importance and redundancy avoidance. Improving text similarity measure is found to be an important driver of any future advances in extractive summarization (Mehta & Prasenjit, 2018). To measure the similarity between sentences (or query and sentences), cluster merging (Section 4.1) and several other standard NLP tasks were performed, for example, sentence splitting, and so forth. This is followed with the semantic representation of each sentence by applying SRL (Section 5.1) with the goal of classifying semantic frames and associated arguments. An example of highly semantically related sentences highlighted in (Mohamed & Oussalah, 2019) is re-used here (Example 2) for illustration purposes only.

Example 2. S_1: FIFA was accused of corruption.
S_2: FIFA was officially investigated for corruption.

The application of SRL enables the identification of sentence predicate verbs. In the above example, every sentence has one predicate verb (semantic frame), that is, *accuse* in S_1 and *investigate* in S_2. Arguments of each semantic frame are grouped according to their semantic role to the main verb. Table 3 shows that the semantic parsing identified three common arguments (i.e., A1, A2, & AM-MNR) in the two sentences.

For simplicity, assume that we have two sentences S_1 and S_2, and each has a single semantic frame, that is, f_1 and f_2. Next, let $R_1 = \{r_1, r_2, \ldots, r_k\}$ and $R_2 = \{r_1, r_2, \ldots, r_l\}$ be the semantic roles linked with the semantic frames f_1 and f_2, respectively, where k, and l show the corresponding argument numbers. To compute the similarity between S_1 and S_2, we pair the shared semantic roles, $R_c = \{r_1, r_2, \ldots, r_p\}$, found in both sentences and exclude all other frame roles from similarity measurement. This follows the conjuncture that sentence-to-sentence semantic similarity can be more accurately measured by comparing the arguments of matching semantic frames. Having detected all common semantic roles, we next construct role-term vectors, $TV = \{WV_{1i}, WV_{2i}, \ldots, WV_{pi}\}$, corresponding to paired common semantic roles R_c as in Table 4.

14680394, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13462 by Test, Wiley Online Library on [02/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

The table lists pre-processed normalized argument terms of the shared semantic frames. The argument terms are then mapped to corresponding Wikipedia concepts with the help of the pre-built inverted index file (see Section 5.2). In other words, the role-argument terms (Table 4) are translated to a table of Wikipedia concept vectors. For example, if WV_{ij} denotes the argument terms of semantic role i in sentence j, we create CV_{ij}, the weighed vector of Wikipedia concepts from WV_{ij}. For illustration purpose, Table 5 lists the first five Wikipedia concepts associated with the argument term *investigated* along with their unique Wikipedia IDs and TF-IDF scores.

Finally, the semantic similarity between S_1 and S_2 is calculated using these fill-in Wikipedia concepts. Formally, the argument terms of all semantic roles found in both sentences, $(r_1,\ldots,r_p,\ p = number\ of\ shared\ roles)$ are converted to their semantically linked Wikipedia concept vectors, $\{CV_{11},\ldots,CV_{i1}\}$ and $\{CV_{12},\ldots,CV_{i2}\}$. Then, the semantic similarity between S_1 and S_2 is computed as their average role similarities (RSim). This can be mathematically expressed as in Equation (4).

$$Sim_{srl-esa}(S_1,S_2) = \frac{1}{p}\sum_{i=1}^{p} RSim(CV_{1i},CV_{2i}) \tag{4}$$

The entity $RSim(CV_{1i},CV_{2i})$ is computed using individual concepts representing the original argument terms as formulated in (5). In Equation (5), wc_{j1} represents the $tf*idf$ weight of j^{th} term with respect to its corresponding concept from argument role i of sentence 1, while wc_{j2} is the $tf*idf$ weight of term j with respect to its corresponding concept from argument role i of sentence 2.

$$RSim(CV_{1i},CV_{2i}) = \frac{\sum_{j=1} wc_{j1} * wc_{j2}}{\sqrt{\sum_{j=1} wc_{j1}^2}\sqrt{\sum_{j=1} wc_{j2}^2}} \tag{5}$$

Figure 5 provides a high-level demonstration of the SRL-ESA based approach for estimating the semantic similarity between sentences S_1 and S_2. The figure distinguishes four distinct stages with the assumption of three shared roles. In the first stage, the semantic parsing using

TABLE 3 Tokenization and semantic parsing of the sentences in Example 2.

S_1 predicates and semantic arguments			S_2 predicates and semantic arguments		
Terms	Predicates	Role tags	Terms	Predicates	Role tags
FIFA	–	A1	FIFA	–	A1
was	–	0	was	–	0
accused	*Accused*	V	officially	–	AM-MNR
of	–	B-A2	investigated	*Investigated*	S-V
corruption	–	E-A2	for	–	B-A2
–	–	–	corruption	–	E-A2

TABLE 4 Role-term vectors: shared semantic frames/roles with associated term vectors.

Role (Arg.) label	S_1 argument terms (WV_{i1})	S_2 argument terms (WV_{i2})
V	Accuse	Investigate
A1	FIFA	FIFA
A2	Corruption	Corruption

TABLE 5 The first five Wikipedia concepts for the argument term: *Investigated*.

Wikipedia ID#	Concepts	TF*IDF weight
3634121	Investigative reporters and editors	0.5345
11917620	United States house energy subcommittee on oversight and investigations	0.4757
11676740	Crime & investigation network	0.4589
43032911	Special investigations	0.4502
5236980	Criminal investigation	0.4202

14680394, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13462 by Test, Wiley Online Library on [02/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 6 Sentence ranking features for the proposed semantic-based Qf-MDS.

Query-dependent features		Query-independent features	
Feature	Notation	Feature	Notation
Query similarity	QS	Sentence centrality	SC
Title similarity	TS	Position	P
Query cosine similarity	QCS	Centroid	C
Named entity overlap	NEO	Sentence length	L
Query terms overlap	QTO	–	–

semantic role labelling is applied to the two inputted sentences generating semantically tagged/parsed sentences. In the second phase, we identify the sentence predicate verbs and their associated semantic role-sets. The third step selects the arguments of common semantic roles and discards unshared roles. Next, grouped argument terms are translated to the corresponding Wikipedia concepts in the last stage. Finally, the actual similarity between the two sentences is computed from the representative Wikipedia concepts using cosine similarity measure.

6.2 | Assessing sentence importance and query-relevance

The proposed Qf-MDS framework is underpinned with a combination of nine features to assess the importance and query-relevance of document sentences for summary inclusion. As shown in Table 6, the summarizer is built on two types of scoring features: query-dependent and query-independent. The features are selected to ensure the consideration of both query-relevance and coverage of the concepts expressed in the documents' sentences before they are scored, ranked, and selected for summary inclusion.

6.2.1 | Query-dependent features

The five different features presented in the first column of Table 6 are used to determine query-relevance and capture semantic and lexical similarity information between the cluster queries and sentences. The *Query Cosine Similarity* and *Query Terms Overlap* features are specifically used to design a baseline summarizer in addition to being part of the overall features. The two important and core query-relevance scoring features built on the SRL-ESA similarity measure (see Section 6.1) are the *Query Similarity* and *Title Similarity*. These five query-dependent features are briefly defined below.

1. *Query Similarity (QS) (or equivalently, Query relevance (QR))* is the heart of any query-focused summarization task. It is measured in terms of sentences' semantic relatedness with the given query by defining the semantic association between knowledge-based concept vectors of the query (Q) and cluster sentences. Query-focused summarization can be thought as a question answering task where the query is the question that expresses the user's information need and the summary is the answer. For example, the DUC2006's D0610A cluster is accompanied by this query *"What are the advantages and disadvantages of home schooling? Is the trend growing or declining?"*. In the context of this paper, the Query Similarity (expression 6) is built on the SRL-ESA based similarity measure in Equation (4) with the Q and S_i representing the query and sentence i to be compared, respectively.

$$QS(Q,S_i) = Sim_{srl-esa}(Q,S_i) \tag{6}$$

2. *Title similarity (TS)*: Each cluster of the used experimental data (i.e., DUC2006) is provided with a title that summarizes the topic and concepts discussed in its body. Related to this, one may deduce that semantically related sentences to the cluster title are good candidates for summary inclusion. The TS feature is designed to capture this aspect of the association by computing the semantic relatedness between commonsense Wikipedia concepts interpreted from cluster titles and sentences. Unlike cluster queries and sentences, titles mainly comprise of noun phrases with no semantic frames and predicate verbs, e.g., *steroid use among female athletes*. For that reason, we computed the *title similarity* from Wikipedia concepts without semantic parsing (i.e., semantic role labelling is not applied to cluster titles). The TS feature for sentence s_i is computed as per expression (7), where T is the title, and $Sim_{esa}(T,S_i)$ is achieved with Equation (3).

$$TS(T,S_i) = Sim_{esa}(T,S_i) \tag{7}$$

14680394, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13462 by Test, Wiley Online Library on [02/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3. *Query term overlap (QTO)*: This feature calculates the lexical overlap between cluster query and sentences. It is used as a contributing query-relevance feature as well as a baseline feature with the objective of giving preference to cluster sentences with high lexical co-occurrence with the user question (query). Assuming $|Q|$ to be the number of query words and $|S_i|$ be the cardinality of the terms in the i^{th} sentence, the QTO feature can be computed as per Equation (8).

$$QTO(Q,S_i) = \frac{|Q \cap S_i|}{(|Q \cup S_i|) - (|Q \cap S_i|)} \tag{8}$$

4. *Query cosine similarity*: The query cosine similarity (QCS) feature is employed to compute the cosine similarity between the cluster query and its sentences. It is also used to complement the QTO feature in developing a baseline summariser. If we assume $\vec{Q}$ and $\vec{S_i}$ to be word vectors of the cluster query and sentences, the QCS feature is formulated as per Equation (9):

$$QCS(Q,S_i) = \frac{\sum_{i=1} q_i * s_i}{\sqrt{\sum_{i=1} q_i^2}\sqrt{\sum_{i=1} s_i^2}} \tag{9}$$

The notations s_i and q_i in the above equation represent the the TF-IDF weights for words w_{is} in the sentence and query in order.

5. *Named entity overlap (NEO)*: A named-entity is a term used in NLP to refer to the proper names of real-world objects such as people, places, organizations, and so forth. The named entity similarity between cluster queries and sentences is evaluated based on their overlap. Intuitively, the existence of named entities in a user query suggests that its answer may be found in document sentences containing similar named entities. Therefore, we used the named entity overlap measure in Equation (10) to boost the similarity measures of queries and sentences sharing lexically identical entities. The feature is built on the concept of Jaccard similarity measure, that is, if we let QE represent the named-entities in query Q and SE be the set of named-entities present in sentence S, then the NEO feature can be quantified as per Equation (10).

$$NEO(Q,S_i) = \frac{QE \cap SE}{QE \cup SE} \tag{10}$$

6.2.2 | Query-independent features

This category of features separately determines the importance of cluster sentences and includes the *centrality*, *centroid*, *length* and the *position* of each sentence. The centrality and centroid features are used to capture the semantic coverage of a cluster sentence, as briefly defined below.

1. *Sentence semantic content (SSC)*: It is widely acknowledged in related literature that the use of query-relevance as the only scoring feature cannot produce the best summary to meet the information need of a typical user question (Canhasi & Kononenko, 2014). Two often interchangeably used features for addressing this shortfall in text summarization are the centrality and coverage (Mohamed & Oussalah, 2015). In this work, we model the centrality feature using two parameters: SSC and Centroid. The SSC measure corresponds to the amount of semantic information present in a given cluster sentence. Formally, the SSC score of a given sentence s_i (Equation 11) is computed by averaging the similarity score between current and other cluster sentences without including sentences from the same document D_i.

$$SSC(s_i) = \frac{1}{|C| - |D_i| - 1} \sum_{s_j \in C - D_i} Sim(s_i, s_j) \tag{11}$$

In Equation (11), the $|C|$ denotes the number of cluster sentences; $|D_i|$ represents the number of sentences in the document from which the current one (s_i) is taken; $C - D_i$ indicates the sentences in the cluster excluding those in document D_i.

2. *Centroid*: This is the second sentence centrality component that we used in this work. It is an empirically determined set of words that can statistically represent a cluster of related documents about the same topic (Radev, 2004). In other words, the sentence centroid score, denoted as $C_i(s_i)$, is computed by summing the centroid scores of individual words, C_{w_i}, in that sentence as formulated in expression (12).

$$C(s_i) = \sum C_{w_i},\ where\ C_{w_i} = TF_{w_i} * IDF_{w_i} \tag{12}$$

3. *Sentence length*: The sentence length feature ($L(s_i)$) is the number of words ($|s_i|$) in it. Most summarization datasets come with a requirement that the generated summary should not exceed a given number of words. With such criteria, sentence length became an important feature.

14680394, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13462 by Test, Wiley Online Library on [02/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

For example, if the generated summary contains short sentences, the conveyed information may be limited, undermining the summary quality. On the other hand, with longer sentences, one will quickly reach the maximum permitted number of words. For that reason, we set a preference for sentences whose lengths are up to 10 words only, to achieve a trade-off of maximum coverage in the summary.

$$L(s_i) = |s_i| \tag{13}$$

4. *Sentence position*: The location of a sentence in a document is used to indicate its importance. This is particularly the case in the news domain where documents are structured such that the start or end sentences convey some core important information. This work uses the position feature as a sentence importance scoring feature since the used experimental dataset, the DUC2006, belongs to the news domain. The position of a sentence ($P(s_i)$) is formulated as the reciprocal of the cluster size ($|C|$). In this way, feature values are assigned in a decreasing order with the first sentence receiving the highest score and with the last one getting the least score.

$$P(s_i) = \frac{1}{|C|} \tag{14}$$

6.3 | Sentence ranking and summary extraction

The principle aim and the heart of the proposed semantic diversity feature based query-focused summarization approach is to score and rank cluster sentences for summary extraction. In other words, a composite scoring function (Equation 15) is first applied to rank sentences after which they are re-ranked with *MMR* algorithm (see Section 4.2). The scoring function is designed so that the sentence score is calculated as a linear combination of its weighted feature scores (cf. Table 6).

$$Score(s_i) = \sum_{j=1}^{n} w_{f_j} * f_j(s_i) \tag{15}$$

In the above equation, s_i shows the i^{th} sentence, w_{f_j} denotes the weight associated with the sentence feature, $f_j(s_i)$, while n represents the number of combined features. We have searched for the best feature weights by experimenting with various weight values from within a given range (see Section 7.2).

Following the ranking of the cluster sentences as per Equation (15), we have applied the *MMR* algorithm underpinned with our SRL-ESA based similarity measure to re-score and re-rank sentences for the purpose of preventing redundancy and achieving diversity in the produced summary. In this work, we have used a slightly modified version of the *MMR* in order to improve summary diversity with enhanced semantic similarity identification as per Equation (16). This is achieved by replacing Sim_2 with two metrics: the similarity of each candidate summary sentence (s_i) with already selected summary (*Sum*), and a penalty factor that prevents selecting sentences from documents that the algorithm already picked from; $f(s_i, D_i, Sum) = \frac{1}{|D_i|}\sum[s_i \in Sum]$.

$$S_{MMR}(s_i) = \lambda Score(s_i) - (1-\lambda)[Sim(s_i, Sum) + f(s_i, D_i, Sum)] \tag{16}$$

where the $Score(s_i)$ is the accumulative sentence score computed as per Equation (15), and s_i represents sentence i previously included in the summary *Sum* from document D_i. The extracted summary sentences are finally organized according to the order they appeared in the source cluster document.

7 | EXPERIMENTS

7.1 | Evaluation dataset and measures

To evaluate the proposed summarization system, we conducted experiments on a dataset constructed from the DUC2006 corpus (http://duc.nist.gov/data.html), a standard human annotated dataset specifically created for the performance assessment of complex real-world user-oriented multi-document summarization systems. The dataset contains 50 clusters of about 25 documents each along with user queries and corresponding gold summaries with the objective of extracting a summary answer of about 250 words from each cluster. Table 7 provides a brief description of the evaluation dataset. To quantify the performance of the proposed summarizer against baselines and related studies, we used the Recall Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004), which is the most widely automated evaluation tool used in text summarization.

14680394, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13462 by Test, Wiley Online Library on [02/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 7 Statistical description of the experimental dataset, DUC2006.

Statistic	Values (s)
Data source	AQUAINT
Numer of clusters	50 clusters
Number of documents per cluster	25 documents
Length of summaries	250 words
Cluster size range in sentences	165 to 1349 sentences
Average cluster size in sentences	717 sentences
Cluster size range in words	4312 to 23,285 words
Average cluster size in words	13,442 words
Summarization task	Query-oriented / Complex QA

Typically, ROUGE (Equation 17) determines the quality of auto-generated summaries by comparing their textual content to a human generated reference summary and computing a group of ROUGE measures including ROUGE-1, ROUGE-2, and ROUGE-SU4.

$$ROUGE-N = \frac{\sum\limits_{S \in Ref\ Summ,} \sum\limits_{gram_n \in S} Count_{match}(gram_n)}{\sum\limits_{S \in Ref\ Summ,} \sum\limits_{gram_n \in S} Count(gram_n)} \tag{17}$$

In the above equation, N represents the length of the n-gram ($gram_n$), $Count(gram_n)$ corresponds to the number of n-grams ($gram_n$) in the reference summary while $Count_{match}(gram_n)$ denotes the maximum number of n-grams co-occurring in the auto-generated system summary and the collection of reference summaries (RefSumm). A $gram_n$ refers to a chain of n words, for instance, a two-word sequence is called a bigram. In the context of this paper, we will be using three selected ROUGE measures, namely ROUGE-N (N = 1, 2) and ROUGE-SU4. These measures were empirically found to work well for query-focused summarization tasks (Lin, 2004)

7.2 | Empirical results and discussion

In the first step of competing information redundancy in document clusters, we have applied the iterative merging algorithm described in Section 4.1 for the purpose of unifying related cluster documents into a single document while filtering out redundant sentences. This produced a single document for each cluster in which all highly similar sentences are reduced to a single representative sentence. Figure 6 shows DUC2006 cluster sizes before and after merging. As shown, the original DUC2006 document cluster sizes range from about 160 sentences to over 1300. One can also deduce from this illustration that larger document clusters contain more redundant information than the smaller document sets as they constitute more sentences. The application of this similarity filtering algorithm to the dataset reduced its cluster sizes to remove redundancy, while also speeding up subsequent data processing.

We run a statistical test (t-test) to check whether the merging process has significantly reduced the cluster sizes, hence substantially removing redundancy and maintaining the diversity among summary sentences to be generated. The statistical test was based on the following hypothesis: $H_0 : \mu_a - \mu_b > = 0$; $H_A : \mu_a - \mu_b < 0$; where μ_b and μ_a respectively show the mean cluster sizes before and after merging. Using 95% confidence interval, we found evidence that the merging and redundancy removal algorithm has significantly reduced the document cluster sizes ($t(49) = -10.88$, $p < 0.0001$) thereby removing information repetition and maintaining diversity in summaries, which is one primary challenge with multi-document summarization.

Next, we conducted a set of experiments to extract cluster summaries, truncated to 250 words, using various sentence scoring feature combinations. We tested the performances of the aggregated features in terms of the three used ROUGE measures (ROUGE-1, ROUGE-2, ROUGE-SU4) and their results are given in Table 8. Notably, all combinations of the sentence scoring methods include the query (QS) and title (TS) similarity features both which are built on developed SRL-ESA similarity measures. The fact that the aggregate of the two features achieves almost the same performance as all the combined scoring methods, indicates that Wikipedia concepts translated from the same semantic arguments effectively capture the sentence similarity. In this initial experiment, all features are linearly combined without applying any weighting mechanism. Best results among these unweighted feature combinations are bold-highlighted in Table 8. For comparison purposes, we have also designed a simple baseline summarizer that uses two query-dependent features, namely, QCS and QTO. The ROUGE results of this baseline summarizer are also given at the end of the same table. Please note that all results are rounded to three significant figures.

14680394, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13462 by Test, Wiley Online Library on [02/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

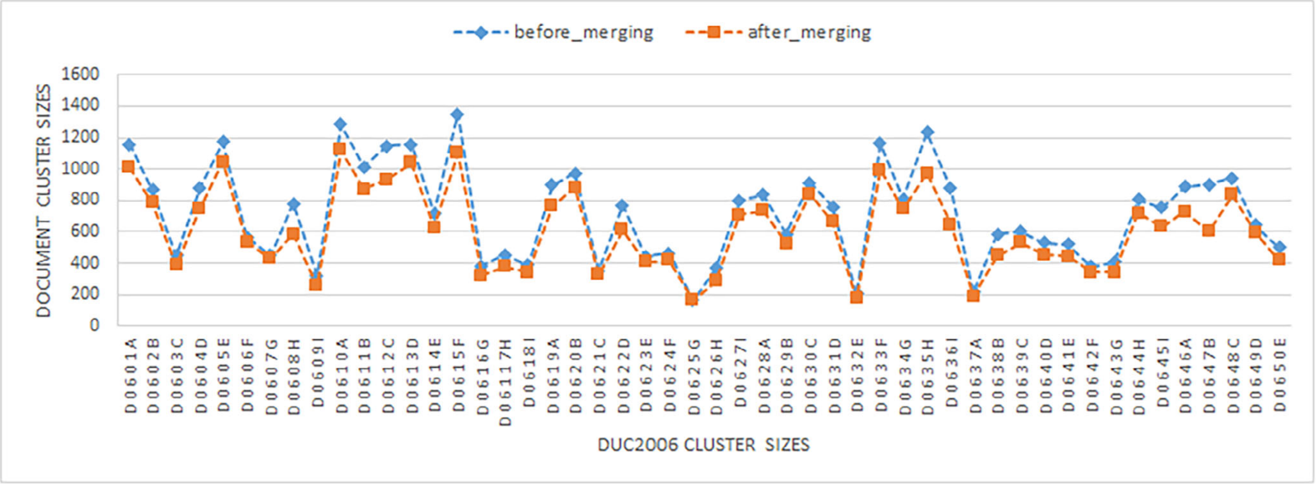


FIGURE 6 DUC2006 cluster sizes (no of sentences) before and after merging.

TABLE 8 Comparing results of the proposed semantic diversity feature based query-focused summarizer using different unweighted feature combinations: average recall of the three selected ROUGE measures at 95% confidence interval.

QS	TS	L	SC	NEO	C	P	QCS	QTO	ROUGE-1	ROUGE-2	ROUGE-SU4
✓	✓	✓							0.410	0.089	0.148
✓	✓	✓	✓	✓					0.412	**0.093**	**0.150**
✓	✓	✓	✓	✓	✓				0.413	0.091	0.149
✓	✓	✓	✓	✓	✓	✓			0.413	0.090	0.148
✓	✓	✓	✓	✓	✓	✓	✓		0.413	0.090	0.148
✓	✓	✓	✓	✓	✓	✓	✓	✓	**0.417**	0.092	0.115
							✓	✓	0.358	0.056	0.117

In addition, we examined the impact of weighting the sentence scoring features on the effectiveness of the similarity measures and the overall summarizer. To this end, we iteratively searched and obtained optimum weight values for all scoring features in a specified interval of [1-5][‡]. The feature weighting coefficients were semi-manually optimized to maximize the ROUGE recall scores for the three measures on the DUC2006 dataset by comparing human created and auto-produced system summaries. Tables 9–11 provide the summarizer's overall performance results for the selected three ROUGE measures based on the combination of weighted features. The "Average" column represents the mean ROUGE scores, while "Best" and "Worst" correspond to best and worst recorded scores of the respective measure. Overall, the results in the tables show that weighing features slightly improves the performance of the proposed summarizer compared to the systems based on unweighted features (see Table 8).

Notably, there is a considerable variation in performance among the different dataset clusters, which shows that some cluster summaries generated by the proposed system have more content overlap with the human summaries (cf. "Best") than others (cf. "Worst"). The columm, denoted by "*Diff.*" in the table shows that approximate average deviations between best and worst performing document clusters are 13%, 8%, 9% for the ROUGE-1, ROUGE-2, and ROUGE-SU4 measures in order. The differences in performance among the cluster summaries can be primarily attributed to the large document sizes (i.e., number of sentences) of the merged related document sets. It could also be due to the high compression rates needed to summarize such large sentence pools to a few summary sentences. As previously mentioned, the DUC2006 dataset contains document sets ranging from 160 to just over 1300 sentences. Intuitively, one would expect a better content match in a summary selected from the 160-sentence cluster as compared to one taken from 1300-sentence cluster.

7.2.1 | Comparison with baselines and related works

To further investigate the performance of the semantic diversity feature based query-focused summarization system, we carried out a performance comparison with a range of relevant baselines and related state-of-the-art published works. These comparators include our own baseline,

14680394, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13462 by Test, Wiley Online Library on [02/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the average score of all DUC2006 participating systems, the highest ranked pertinent DUC2006 system, and several other benchmark methods, all tested on the same dataset (DUC2006) as other compared benchmark methods. This performance comparison of the semantic diversity feature based summarizer and other methods is given in Table 12. The related studies used to compare our results were selected based on their methodological similarities to our work (e.g., sentence scoring features, use of semantic methods, diversity consideration, etc.) and provided the same evaluation dataset (DUC2006) was used as well. The ROUGE scores presented in the table are average recall values at 95% confidence interval and the ranking of each comparator is denoted by the number in parenthesis following its ROUGE score. Below are brief descriptions of the used benchmark methods and related studies.

TABLE 9 ROUGE-1 results of the proposed semantic diversity feature based query-focused summarisation using weighed features: average measure scores at 95% confidence interval.

Measure	Average	Best	Worst	Diff.
Recall	0.4182	0.5035	0.3551	0.1484
Precision	0.3865	0.4488	0.3261	0.1227
F-measure	0.4014	0.4671	0.3404	0.1267

TABLE 10 ROUGE-2 results of the proposed semantic diversity feature based query-focused summarisation using weighed features: average measure scores at 95% confidence interval.

Measure	Average	Best	Worst	Diff.
Recall	0.092	0.1331	0.0474	0.0857
Precision	0.0854	0.1227	0.0454	0.0773
F-measure	0.0885	0.1277	0.0464	0.0813

TABLE 11 ROUGE-SU4 results of the proposed semantic diversity feature based query-focused summarisation using weighed features: average measure scores at 95% confidence interval.

Measure	Average	Best	Worst	Diff.
Recall	0.1519	0.2083	0.1089	0.0994
Precision	0.1404	0.1788	0.098	0.0808
F-measure	0.1458	0.1925	0.1032	0.0893

TABLE 12 Comparison of our summarizer's ROUGE results with baseline methods and related works–all figures are rounded to three significant figures.

Summarizer	ROUGE-1	ROUGE-2	ROUGE-SU4
Baseline	0.358 (10)	0.056 (10)	0.117 (9)
AVG-DUC2006	0.3795 (9)	0.075 (8)	0.132 (8)
Best DUC2006 (Jagarlamudi et al., 2006)	0.380 (8)	0.095 (2)	0.155 (2)
Regression (Ouyang et al., 2011)	–	0.093 (3)	0.149 (4)
wAASum (Canhasi & Kononenko, 2014)	**0.424** (1)	0.092 (4)	**0.167** (1)
RDRP_AP (Cai & Li, 2012)	0.396 (7)	0.090 (5)	0.139 (7)
PRCN (Luo et al., 2013)	0.409 (4)	0.092 (4)	0.144 (6)
MCLR (Alguliev et al., 2012)	0.398 (6)	0.085 (7)	–
EBSS (Bidoki et al., 2020)	0.405 (5)	**0.113** (1)	–
CAiRE-COVID (Su et al., 2020)	0.345 (11)	0.065 (9)	0.112 (10)
SbQf-HypSum (Mohamed & Oussalah, 2015)	0.412 (3)	0.088 (6)	0.147 (5)
SDbQfSum	0.418 (2)	0.092 (4)	0.152 (3)

14680394, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13462 by Test, Wiley Online Library on [02/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

14680394, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13462 by Test, Wiley Online Library on [02/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1. **Baseline** is the baseline summarizer we implemented using two query-dependent features, namely Query Cosine Similarity (QCS) and Query Term Overlap (QTO) as in Equations (9) and (8) respectively.
2. **Best DUC2006** is the best participating system in DUC2006 query-focused summarization competition (Jagarlamudi et al., 2006). The summarizer uses two features, a query-dependent and a query-independent sentence importance feature, to score each sentence for summary inclusion
3. **Regression** is a query-based text summarization approach built on regression models (Ouyang et al., 2011). Specifically, the authors used support vector regression to score and identify important query-relevant sentences using a set of features including word, semantic and named entity overlap
4. **wAASum** is a system built on the concept of matrix factorization and weighted archetypal analysis to generate summaries that meet user information needs represented in the form of a query (Canhasi & Kononenko, 2014). In particular, a three-element graph representation consisting of documents, sentences, and words is used to model the summarizer.
5. **RDRP_AP** is a query-focused summarization model based on manifold-ranking and improved with mutual reinforcement between sentences and theme clusters (Cai & Li, 2012). In this approach, the summarization is achieved by building a weighted graph of cluster sentences and queries.
6. **PRCN** is a query-oriented summarization method that simultaneously optimises query-relevance, topic coverage and information novelty (Luo et al., 2013). The resulting method called Probabilistic-modeling Relevance, Coverage, and Novelty (PRCN) ranks sentences for summary inclusion based on features that represent the three aforementioned factors.
7. **MCLR** is a multi-document summarization framework developed with the objective of concurrently optimising the coverage and redundancy (Alguliev et al., 2012). The system is modelled by aggregating weighted content coverage and redundancy features in order to produce non-redundant and central summary sentences.
8. **EBSS** is a semantic-based multi-document summarization system constructed on the combination of statistical, graph-based, and machine learning methods while also using word vectors for semantic representation (Bidoki et al., 2020). In addition, the authors claimed their proposal to be language independent as well.
9. **CAiRE-COVID** is a question answering (QA) and multi-document summarization system developed in participation of Kaggle COVID-19 Open Research Dataset Challenge[§]. The proposal's main aim is to retrieve highly relevant answers from the numerous recent research publications on COVID-19 and summarizing the returned question related answers. It won one of the 10 tasks of the relevant Kaggle challenge.
10. **SbQf-HypSum** is a hybrid summarization system built on WordNet Taxonomy, and enriched with Categorial Variation Database (CatVar) and Morphosemantic Links to improve query similarity and intra-sentences similarities (Mohamed & Oussalah, 2015). It also incorporates Wikipedia-based Normalized Google Distance algorithm to account for named entity semantic relatedness.
11. **SDbQfSum** is the semantic diversity feature based query-focused summarization system proposed in this work.

Table 12 provides the comparison of the ROUGE results between the proposed summarizer and those of benchmark methods and related state-of-the-art (SoA) query-focused summarization approaches. We can observe that this summarizer outperforms most SoA comparators with the exception of a few summarization systems such as the work of Canhasi et al. (2014), Ouyang et al. (2011), and Bidoki et al. (2020) which slightly exceeded our performance in one or two ROUGE measures. The gap in values between the different ROUGE measures, as in Table 12, is due to the disparities in content overlap between the auto-generated and system summaries as captured in these metrics. For example, ROUGE-2 measures the ratio of the number of 2-grams (two-word sequences) in the reference summary that is also extracted as part of the system summary while ROUGE–1 computes the matching 1-grams (single words). For that reason, the ROUGE-2 scores reported in Table 12 are consistently and significantly lower than the corresponding ROUGE–1 values because there are less bigrams present in both the extracted and the reference summaries compared to the corresponding unigram overlaps. Of the comparators in Table 12, the most closely related ones in terms of implementation, considered factors, features, and information source include (Luo et al., 2013; Ouyang et al., 2011). For instance, Luo et al. (2013) exploit similar factors considered in our work, for example, coverage, relevance, and diversity; all modelled using very closely related features. Although, our approach seems to be under performing against one or two comparators in some ROUGE measures rank-wise, the actual differences in performance are not significant. For example, the difference between the best ROUGE-1 and ROUGE-SU4 scoring system (Canhasi & Kononenko, 2014) and ours are actually 0.006 and 0.015 respectively while that of the best performing ROUGE-2 (Jagarlamudi et al., 2006) is 0.021.

Figure 7 includes visual illustration of the comparative evaluation presented in Table 12, but shifted by the baseline scores to improve readability and clarity of their differences. The negative ROUGE-1 and ROUGE-SU4 values of the CAiRE-COVID (Su et al., 2020) indicate that these scores are below the corresponding baseline values. We think that the better performing systems benefit from various factors including the level of implementation, supervision, and granularity of the used text content, as compared to our unsupervised algorithmically simpler approach. For example, the slightly superior work of Canhasi et al. (2014) in ROUGE-1 and ROUGE-SU4 are based on methodologically more complex algebraic methods (e.g., clustering and matrix factorization) and a multi-layered graph representations. They also benefit from the detailed granularity where they use separate but interlinked term, sentence and document similarity graphs thus using more extended semantic information. Other

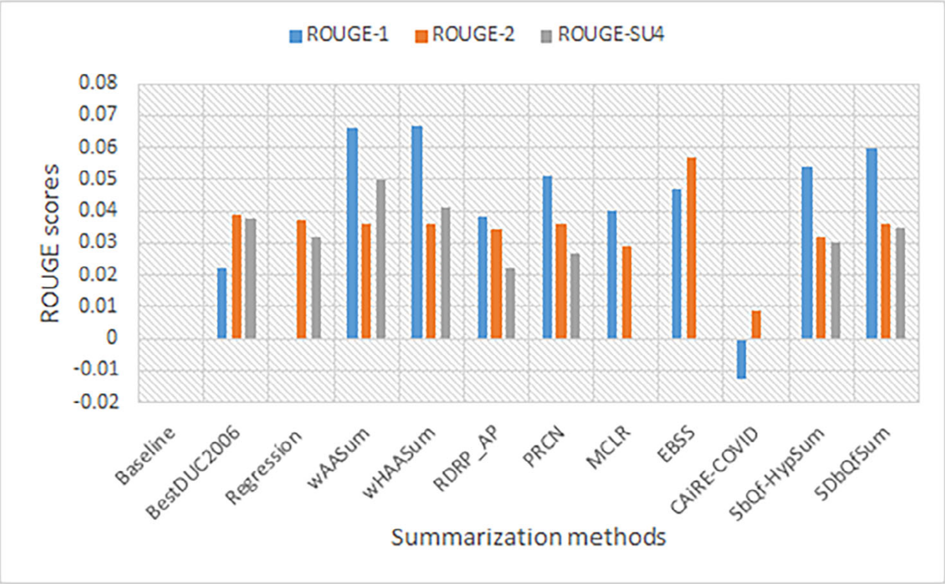


FIGURE 7 Comparison of our system performance with baseline methods and related works.

competitive methods in ROUGE-2 (Bidoki et al., 2020; Ouyang et al., 2011) are built on supervised approaches that learn and utilize human annotated dataset which gives them an advantage over the unsupervised model in this work. Overall, the use of feature-based scoring functions underpinned by effective semantic representation techniques and Wikipedia's rich conceptual commonsense knowledge achieved considerable improvements even though our results are outperformed by one or two state-of-the-art related works.

8 | SUMMARY AND CONCLUSION

We presented a summarization framework for extracting a real-world complex answer summary from a set of related documents. The proposal is built on the principle of extracting a meaningful sentence importance and query-relevance features, which capture the underlying semantic structure beyond the lexical meaning. Extracted features are then used to model the relevance, centrality, and diversity of sentences with the final objective of scoring, ranking, and extracting a few cluster representative sentences as a summary. Scored sentences are then re-ranked using *MMR*algorithm with the aim of preventing redundancy and achieving diversity in the produced summary. We have carried out a set of evaluation experiments starting with the application of a custom cluster merging algorithm for reducing the amount of overlapping arising from the large set of documents assigned to the same topic. A statistical test has shown that the reduction achieved with the merging algorithm significantly reduces the information redundancy while maintaining topic diversity across documents. We have also looked at the effect of weighing sentence importance and relevance features and found that weighting sentence features yields a slightly better performance compared to aggregated unweighted features. To further consolidate our evaluation, we compared our results with a range of other published works on the task of query-focused summarization based on the same dataset. Through this empirical evaluation, we found that the proposed summarization improves the quality of extracted summaries in one or more ROUGE measures as compared to almost all other related state-of-the-art summarization systems.

In the future, we will investigate extending this work by employing neural-based algorithms in order to more effectively analyse the importance of sentence features and extract additional text semantic information. We also intend to apply the proposed semantic representation techniques and features to the task of product and service review summarization.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Muhidin Mohamed https://orcid.org/0000-0002-8449-5818
Mourad Oussalah https://orcid.org/0000-0002-4422-8723
Victor Chang https://orcid.org/0000-0002-8012-5852

14680394, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13462 by Test, Wiley Online Library on [02/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ENDNOTES

[*] http://barbar.cs.lth.se:8081/parse

[†] http://lucene.apache.org.

[‡] Best feature weights were determined as 5.0, 3.0, 5.0, 2.0, 1.0, 1.5, 1.0, 1.0, 1.0 for QS, TS, SC, NEO, P, C, QCS, QTO and L respectively, as tuned from numbers in the interval between 1 and 5.

[§] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

REFERENCES

Abdi, A., Idris, N., Alguliyev, R. M., & Aliguliyev, R. M. (2017). Query-based multi-documents summarization using linguistic knowledge and content word expansion. *Soft Computing*, *21*(7), 1785–1801.

Alguliev, R. M., Aliguliyev, R. M., & Hajirahimova, M. S. (2012). GenDocSum+ MCLR: Generic document summarization based on maximum coverage and less redundancy. *Expert Systems with Applications*, *39*(16), 12460–12473.

Alguliev, R. M., Aliguliyev, R. M., & Isazade, N. R. (2013). Multiple documents summarization based on evolutionary optimization algorithm. *Expert Systems with Applications*, *40*(5), 1675–1689.

Alguliyev, R. M., Aliguliyev, R. M., Isazade, N. R., Abdi, A., & Idris, N. (2019). COSUM: Text summarization based on clustering and optimization. *Expert Systems*, *36*(1), e12340.

Al-Sabahi, K., Zhang, Z., Long, J., & Alwesabi, K. (2018). An enhanced latent semantic analysis approach for arabic document summarization. *Arabian Journal for Science and Engineering*, *43*(12), 8079–8094.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley framenet project. COLING 1998 volume 1: The 17th international conference on computational linguistics.

Bedi, P. P. S., Bala, M., & Sharma, K. (2022). Extractive summarization using concept-space and keyword phrase. *Expert Systems*, *39*(10), e13110.

Bidoki, M., Moosavi, M. R., & Fakhrahmad, M. (2020). A semantic approach to extractive multi-document summarization: Applying sentence expansion for tuning of conceptual densities. *Information Processing & Management*, *57*(6), 102341.

Binwahlan, M. S., Salim, N., & Suanmali, L. (2010). Fuzzy swarm diversity hybrid model for text summarization. *Information Processing & Management*, *46*(5), 571–588.

Cai, X., & Li, W. (2012). Mutually reinforced manifold-ranking based relevance propagation model for query-focused multi-document summarization. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(5), 1597–1607.

Canhasi, E., & Kononenko, I. (2014). Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization. *Expert Systems with Applications*, *41*(2), 535–543.

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

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, *12*, 2493–2537.

Dong, Y. (2018). A survey on neural network-based summarization methods. arXiv preprint arXiv:1804.04589.

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.

Ferreira, R., Souza Cabral, d. L., Freitas, F., et al. (2014). A multi-document summarization system based on statistics and linguistic treatment. *Expert Systems with Applications*, *41*(13), 5780–5787.

Fillmore, C. J., & Baker, C. F. (2012). A frames approach to semantic analysis. In *The Oxford handbook of linguistic analysis* (pp. 313–339). Oxford University Press.

Gabrilovich, E., & Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, *34*, 443–498.

Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: A survey. *Artificial Intelligence Review*, *47*(1), 1–66.

Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, *28*(3), 245–288.

Gupta, S., & Gupta, S. K. (2019). Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, *121*, 49–65.

He, L., Lee, K., Lewis, M., & Zettlemoyer, L. (2017). Deep semantic role labeling: What works and what's next. Proceedings of the 55th annual meeting of the Association for Computational Linguistics (volume 1: Long papers). 473–483.

Jagarlamudi, J., Pingali, P., & Varma, V. (2006). Query independent sentence scoring approach to duc 2006. Proceeding of document understanding conference (DUC-2006).

Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. Text summarization branches out. 74–81.

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

Luo, W., Zhuang, F., He, Q., & Shi, Z. (2013). Exploiting relevance, coverage, and novelty for query-focused multi-document summarization. *Knowledge-Based Systems*, *46*, 33–42.

Mehta, P., & Prasenjit, M. (2018). Effective aggregation of various summarization techniques. *Information Processing & Management*, *54*(2), 145–158.

Meng, Z., & Shen, H. (2018). Dissimilarity-constrained node attribute coverage diversification for novelty-enhanced top-k search in large attributed networks. *Knowledge-Based Systems*, *150*, 85–94.

Mohamed, M., & Oussalah, M. (2019). SRL-ESA-TextSum: A text summarization approach based on semantic role labeling and explicit semantic analysis. *Information Processing & Management*, *56*(4), 1356–1372.

Mohamed, M. A. (2016). *Automatic text summarisation using linguistic knowledge-based semantics*. PhD thesis. University of Birmingham.

Mohamed, M. A., & Oussalah, M. (2015). Similarity-based query-focused multi-document summarization using crowdsourced and manually-built lexical-semantic resources. In: volume 2 of 2015 IEEE Trustcom/BigDataSE/ISPA. IEEE. 80–87.

Mosa, M. A., Anwar, A. S., & Hamouda, A. (2019). A survey of multiple types of text summarization with their satellite contents based on swarm intelligence optimization algorithms. *Knowledge-Based Systems*, *163*, 518–532.

14680394, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13462 by Test, Wiley Online Library on [02/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Murray, G., Renals, S., & Carletta, J. (2005). Extractive summarization of meeting recordings.

Mutlu, B., Sezer, E. A., & Akcayol, M. A. (2019). Multi-document extractive text summarization: A comparative assessment on features. *Knowledge-Based Systems*, *183*, 104848.

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

Ouyang, Y., Li, W., Li, S., & Lu, Q. (2011). Applying regression models to query-focused multi-document summarization. *Information Processing & Management*, *47*(2), 227–237.

Ozsoy, M. G., Alpaslan, F. N., & Cicekli, I. (2011). Text summarization using latent semantic analysis. *Journal of Information Science*, *37*(4), 405–417.

Radev, D. (2004). Centroid-based summarization of multiple documents. *Information Processing & Management*, *40*(6), 919–938.

Sankarasubramaniam, Y., Ramanathan, K., & Ghosh, S. (2014). Text summarization using Wikipedia. *Information Processing & Management*, *50*(3), 443–461.

Su, D., Xu, Y., Yu, T., Siddique, F., Barezi, E., & Fung, P. (2020). CAiRE-COVID: a question answering and query-focused multi-document summarization system for covid-19 scholarly information management. arXiv preprint arXiv:2005.03975.

Verberne, S., Krahmer, E., Wubben, S., & Bosch, V. D. A. (2020). Query-based summarization of discussion threads. *Natural Language Engineering*, *26*(1), 3–29.

Wang, L., Raghavan, H., Cardie, C., & Castelli, V. (2016). Query-focused opinion summarization for user-generated content. arXiv preprint arXiv: 1606.05702.

Zangerle, E., & Bauer, C. (2022). Evaluating recommender systems: Survey and framework. *ACM Computing Surveys (CSUR)*, *55*, 1–38.

Zehlike, M., Yang, K., & Stoyanovich, J. (2022). Fairness in ranking, part ii: Learning-to-rank and recommender systems. *ACM Computing Surveys (CSUR)*, *55*(6), 1–41.

AUTHOR BIOGRAPHIES

Muhidin Mohamed holds a PhD in Natural Language Processing (i.e., Automatic Text Summarization) from the University of Birmingham. He is currently a Teaching Fellow in Business Analytics at the department of Operations and Information Management, Aston University, UK. Dr. Muhidin worked extensively in text mining, information retrieval, data analytics, and information management where he published or contributed to more than 20 international peer-reviewed conference papers and journal articles. He held research visits at several universities including University of Oulu. His current research interests and collaborations include social media data mining, creating AI & NLP resources for low-resourced languages, fake news identification, and AI adoption among SMEs.

Mourad Oussalah (Senior Member, IEEE) received the Ph.D. degree in robotics and artificial intelligence from the University of Paris XII, in 1998. He held research and academic positions with KU Leuven, the City University of London, and Birmingham University, U.K. He is currently a Research Professor with the Faculty of Information Technology and Electrical Engineering, University of Oulu, Finland, and leads Social Mining Research Group. His research interests include information processing, datamining, and text mining. He has published more than 250 technical articles and led several projects in his research areas. He is a fellow of the Royal Statistical Society.

Victor Chang received the Ph.D. degree in computer science from the University of Southampton, Southampton, U.K., in 2013. He is currently a Professor with the Department of Operations and Information Management, Aston Business School, Aston University, Birmingham, U.K. Dr. Chang was the recipient of many awards, including IEEE Service Award 2015, Best Special European Project 2016, Outstanding Young Scienitst 2017, Most Productive AI-based Data Scientist 2010–2019, Top 2%Scientist, 2019, 2020, 2021, Highly Cited Scientist 2021, and so on.

How to cite this article: Mohamed, M., Oussalah, M., & Chang, V. (2023). SDbQfSum: Query-focused summarization framework based on diversity and text semantic analysis. *Expert Systems*, e13462. https://doi.org/10.1111/exsy.13462

14680394, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/exsy.13462 by Test, Wiley Online Library on [02/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License