Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-021-03165-1

ORIGINAL RESEARCH

Unsupervised query-focused multi-document summarization based on transfer learning from sentence embedding models, BM25 model, and maximal marginal relevance criterion

Salima Lamsiyah[1] · Abdelkader El Mahdaouy[2] · Said Ouatik El Alaoui[1] · Bernard Espinasse[3]

Received: 21 October 2020 / Accepted: 13 March 2021

Abstract
Extractive query-focused multi-document summarization (QF-MDS) is the process of automatically generating an informative summary from a collection of documents that answers a pre-given query. Sentence and query representation is a fundamental cornerstone that affects the effectiveness of several QF-MDS methods. Transfer learning using pre-trained word embedding models has shown promising performance in many applications. However, most of these representations do not consider the order and the semantic relationships between words in a sentence, and thus they do not carry the meaning of a full sentence. In this paper, to deal with this issue, we propose to leverage transfer learning from pre-trained sentence embedding models to represent documents' sentences and users' queries using embedding vectors that capture the semantic and the syntactic relationships between their constituents (words, phrases). Furthermore, BM25 and semantic similarity function are linearly combined to retrieve a subset of sentences based on their relevance to the query. Finally, the maximal marginal relevance criterion is applied to re-rank the selected sentences by maintaining query relevance and minimizing redundancy. The proposed method is unsupervised, simple, efficient, and requires no labeled text summarization training data. Experiments are conducted using three standard datasets from the DUC evaluation campaign (DUC'2005–2007). The overall obtained results show that our method outperforms several state-of-the-art systems and achieves comparable results to the best performing systems, including supervised deep learning-based methods.

Keywords Query-focused multi-document summarization · Transfer learning · Sentence embedding models · BM25 model · Semantic similarity · Maximal marginal relevance

1 Introduction

The abundance of textual information on the web raises the need for automatic text summarization (ATS) systems, which aim to produce from one or more documents a fluent summary that synthesize the most relevant information contained in the original documents.

Automatic text summarization systems can be classified based on several factors, including the input, purpose, language, and the output (Nenkova and McKeown 2012). For instance, based on the input factor, we distinguish between single and multi-document summarization. The former generates a summary from one document, while the latter summarizes information from multiple documents. The latter helps users to quickly acquire the relevant information contained in large text collections. Regarding the purpose factor, a number of summarization systems types exist, including generic and query-focused summarization. Generic summarization extracts the important content from the original documents without using any prior knowledge or additional information, whereas query-focused summarization aims to produce a summary reflecting the condensed information that answers the user's information need (expressed by a query). Therefore, as part of our research, we focus on query-focused multi-document summarization.

✉ Salima Lamsiyah
salima.lamsiyah@usmba.ac.ma

[1] Engineering Sciences Laboratory, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco

[2] School of Computer Science (UM6P-CS), Mohammed VI Polytechnic University (UM6P), Ben Guerir, Morocco

[3] LIS UMR CNRS 7020, Aix-Marseille Université/Université de Toulon, Toulon, France

Published online: 25 March 2021

Query-focused multi-document summarization methods are mainly classified into two approaches: ***extractive*** and ***abstractive***. Extractive methods select sentences that are relevant to the input query and subsequently concatenate them to create the final summary, while abstractive methods generate an entirely new summary by concisely paraphrasing the content of the original documents. In this work, we focus on the extractive summarization approach since it usually generates semantically and grammatically coherent sentences.

Query-focused multi-document summarization methods should handle three main issues: (i) how to represent the input document's sentences and users' queries; (ii) how to select relevant sentences to the query; and (iii) how to deal with redundancy.

Documents' sentences and users' queries representation is an important process that affects the effectiveness of query-focused summarization methods. Thus, using a suitable representation for both sentences and queries is extremely important. Bag-of-words (BOW) and word embedding representations have shown promising results in several natural language processing (NLP) tasks, including text summarization (Radev et al. 2004; Jain et al. 2017; Kobayashi et al. 2015; Rossiello et al. 2017). However, these representations do not consider both the ordering of words as well the semantic and syntactic relationships between their constituents (words, phrases). For instance, considering these two sentences: "*The cat ate the mouse*" and "*The mouse ate the cat food*", using either the BOW or word embedding representations, these two sentences will have the same vectors while their meanings are completely different. Hence, we need a more accurate representation of queries and sentences that capture the contextual information and sentences structure.

To overcome the limitations of BOW and word embedding representations for query-focused summarization task, we exploit the potential of transfer learning from the recent pre-trained sentence embedding models for sentences and queries representation. Pre-trained sentence embedding models have recently emerged as a key text representation and achieved impressive performances in a wide variety of NLP tasks (Ethayarajh 2018; Devlin et al. 2019; Cer et al. 2018). These models map variable-length text into dense vectors in a low dimensional vector space, intending to capture the semantic and the syntactic relationships among their constituents. The effectiveness of these models in learning effective contextual representation comes from pre-training them on large amount of unlabeled text data with self-supervised tasks, such as language modeling or filling in missing words, and then transferring the learned knowledge to other downstream tasks. Transfer learning helps in saving time and computational power and benefiting from knowledge learned across other natural language understanding tasks.

Moreover, we use the probabilistic information retrieval model BM25 (BM is an abbreviation of Best Matching) and the semantic similarity from embedding vectors to identify the most representative sentences for the query, from a cluster of documents. Although the BM25 model (Robertson et al. 1995) has shown to be effective in many Information Retrieval tasks; it relies on exact term matching for ranking sentences/documents according to their relevance to the query. Hence, it suffers from the problem of term mismatch (query terms may not occur in relevant sentences). To deal with this issue, we propose to improve sentence retrieval by combining the BM25 and the semantic similarity functions. Semantic similarity is used to measure the semantic closeness of the input query and each sentence in the cluster of documents; it is computed by applying the cosine similarity on their embedding vectors.

Furthermore, we use the Maximal Marginal Relevance method (MMR) (Carbonell and Goldstein 1998) to re-rank the candidate sentences and to produce the final summary. MMR has been widely used in multi-document summarization methods (Lebanoff et al. 2018; Fabbri et al. 2019; Mao et al. 2020) where the summarization task is modeled as the generated summary content should consist of relevant information to the query and minimal similarity among the content. In this work, MMR is applied to deal with redundancy and produce summaries with good information diversity, employing sentence embedding representation.

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

1. We propose an unsupervised query-focused multi-document summarization method based on transfer learning from pre-trained sentence embedding models, BM25 model, and maximal marginal relevance criterion.
2. We explore the potential of transfer learning from three universal pre-trained sentence embedding models. We investigate their effectiveness on the unsupervised query-focused multi-document summarization task using the standard DUC'2005–2007 benchmarks.
3. We combine the BM25 model with the semantic similarity to select a subset of sentences based on their relevance to the query. We show that combining these two functions leads to better performances.
4. We incorporate sentence embedding representation in the maximal marginal relevance method to re-rank the candidate sentences by maintaining query relevance and minimizing redundancy.

To assess the effectiveness of our method, we evaluate its performance with several state-of-the-art methods on DUC'2005–2007 benchmark datasets. The overall results, obtained using ROUGE method, show that our method has achieved promising results on the three datasets; it has

Table 1 Notations used in the paper

Notations	Description
D	Cluster of n documents d
Q	User's query
N	Number of sentences in the cluster D
N_w	Number of sentences containing the word w
S_i	The $i-th$ sentence in the cluster D
$\overrightarrow{Q}$	Embedding vector of the query Q
$\overrightarrow{S_i^D}$	Embedding vector of S_i in the cluster D
x_w^Q	Number of occurrences of word w in Q
$x_w^{S_i}$	Number of occurrences of word w in S_i
l_{S_i}	Length of sentence S_i
l_{avg}	Average of sentence length
$RSV(S_i, Q)$	Retrieval Statue Value of S_i according to the query Q
top-k	Top k ranked sentences in the cluster D
L	Constraint on the summary length

outperformed several unsupervised state-of-the-art systems and achieved comparable results to the best performing systems (CES and Dual-CES) and even outperformed them in terms of R-SU4 evaluation measure. Additionally, it has achieved comparable and sometimes better performance than recent supervised query-focused multi-document summarization systems. Therefore, given its unsupervised nature, simplicity, and effectiveness, we suggest it can be used as a baseline for evaluating query-focused multi-document summarization.

The remainder of this paper is structured as follows: We review the related work in Sect. 2. We describe the proposed method in Sect. 3. Section 4 presents and discusses the experimental results. Finally, we conclude the paper and draw lines for future works in Sect. 5.

The notations we use throughout the paper are summarized in Table 1.

2 Related work

In this section, we first review some state-of-the-art methods for extractive query-focused multi-document summarization. Then, we present the related work on transfer learning from pre-trained sentence embedding models.

2.1 Extractive query-focused multi-document summarization

In recent years, several extractive methods have been introduced for query-focused multi-document summarization. They can be classified into ***supervised*** and ***unsupervised*** methods.

supervised methods fit a model to learn sentences that are relevant to an input query using labeled training data. Supervised machine learning algorithms have been widely used to solve this task, where many extractive query-focused summarization methods consider sentence scoring as a sentence classification or regression problem. For instance, Daumé III and Marcu (2006); Conroy et al. (2005) have exploited hidden Markov and Bayesian statistical models to extract query features to estimate sentences' saliency. Celikyilmaz and Hakkani (2010) have developed HybHSum, an extractive method for query-focused multi-document summarization based on a two-step learning model: a hierarchical probabilistic model used for discovering the topic structures of all sentences, and a regression model for inference. Ouyang et al. (2011) have applied a support vector regression model to rank sentences based on their relevance to a pre-given query where ground truth labels have been generated by computing similarity between sentence and reference summary using several N-gram based methods. Following the success of supervised deep learning models in generic summarization methods (Lebanoff et al. 2018; Fabbri et al. 2019; Eheela and Janet 2020), many researchers have adopted deep learning models to address the query-focused summarization. Valizadeh and Brazdil (2015) have proposed an extractive method that uses an ensemble summarizing system to select the sentences that satisfying actor-object relationships. The method extracts sentences features based on a graph topology and then pass them through a feed forward neural network for sentence selection learning. Cao et al. (2016) have introduced AttSum system that applies convolutional neural networks with attention mechanism to jointly tackle query relevance and sentence saliency ranking tasks. It automatically learns sentences embeddings representations as well as the document cluster. In the same context, Ren et al. (2017) have developed CRSum-SF system that combines a convolutional neural network and a recurrent neural network to jointly learn sentences representations and similarity scores between a sentence and sentences in its context. Similarly, Ren et al. (2018) have introduced SRSum (Sentence Relation-based summarization) system, which uses convolutional neural networks with attention mechanism to select the relevant sentences to the input query and the context. More recently, Sakai et al. (2019) have employed eight textual similarity measures to generate ground truth labels at the sentence level given a reference summary. They have then used these labeled data to train different deep learning models to produce extractive summaries. They have showed that ROUGE-WE2 and ROUGE-SU similarity measures have achieved the best performances. Indeed, supervised methods may produce better quality summaries; however, they require large amounts of labeled training data, limiting their applicability when labeled data are scarce. Therefore, generalizing supervised methods to new

domains and languages remains an open problem. To this end, query-focused multi-document summarization has been addressed mostly using unsupervised methods (Nenkova and McKeown 2011).

Unsupervised methods are mainly based on scoring documents' sentences by combining a set of predefined features. Various studies extract relevant and query-dependent sentences using graph-based approach. For instance, Wan and Xiao (2009) have introduced MultiMR graph-based manifold-ranking method for extractive query-focused multi-document summarization, which considers the within and the cross-document sentence relationships. Canhasi and Kononenko (2014) have proposed wAASum, a novel graph-based method that represents the input documents and query as multi-element graph, and then uses a weighted archetypal analysis factorization method to estimate the importance of sentences based on their relevance to the query. Xiong and Ji (2016) have introduced an hypergraph-based vertex reinforcement ranking framework for extractive query-oriented summarization that integrates several factors for sentence ranking. Van Lierde and Chow (2019a) have proposed a graph-based method for query-focused multi-document summarization that uses a fuzzy hypergraph model to infer topic distributions of sentences. In the same context, Van Lierde and Chow (2019b) have introduced a novel method for extractive query-oriented summarization that applies hypergraph transversals to generate query-relevant summaries. Other researchers have based their works on statistical latent models. Haghighi and Vanderwende (2009) have introduced HierSum, an unsupervised query-focused multi-document summarization method that uses a hierarchical LDA-style model to represent content specificity as a hierarchy of topic vocabulary distributions. Similarly, Shen et al. (2011) have proposed BI-PLSA a variant of the probabilistic latent semantic analysis method that simultaneously clusters and summarizes documents. Besides, Yao et al. (2015) have proposed SpOpt, an unsupervised method for query-focused multi-document summarization based on sparse optimization with decomposable convex objective function where summaries are generated by minimizing the documents reconstruction error and maximizing the dissimilarity between the selected sentences. Wu et al. (2019) have introduced DPRQSum, an unsupervised method for query-focused multi-document summarization that employs dual pattern-enhanced models, the first one is used to generate discriminative and semantic representation for topics while the other is used to model for query-relevance of sentences. Meanwhile, deep learning models have also shown promising results in the context of query-focused summarization task. Wan and Zhang (2014) have developed CTSUM system for query-focused multi-document summarization that incorporates information certainty in summarization process by automatically predicting sentence uncertainty scores. Ma et al. (2016) have developed DocRebuild framework, which reconstructs the documents with summary sentences through a neural document model and generate the summary by extracting sentences that minimize the reconstruction error. The proposed system uses two different text representations Bag-of-Words and the Paragraph Vector where the latter representation yields better performances. Moreover, Zhong et al. (2015) have developed QODE an unsupervised query-focused multi-document summarization method that combines the Restricted Bolzmann Machines and dynamic programming. Yousefi-Azar and Hamey (2017) have proposed an unsupervised method for query-focused summarization based on a stochastic version of deep auto-encoders called the Ensemble Noise Auto-Encoders, which add some noise to the input text representation and then select the relevant sentences from an ensemble of noisy runs. Nevertheless, Feigenblat et al. (2017) have recently introduced CES an unsupervised query-focused multi-document summarization approach based on the Cross-Entropy method. In the same context, Roitman et al. (2020) have proposed Dual-CES, an unsupervised system that employs a two-step dual-cascade optimization approach with saliency-based pseudo-feedback distillation to better handle the trade-off between saliency and focus in summarization. Both CES and Dual-CES systems do not require domain knowledge and achieved the state-of-the-art performances on the three DUC'2005–2007 datasets.

2.2 Transfer learning from sentence embedding models

Recent years have featured a trend towards applying transfer learning techniques in several fields including computer vision and natural language processing. The objective of transfer learning is to improve the learning of the target task by using the knowledge gained from a source task, which helps boost the performance of this target task. In other words, a model is first pre-trained on data-rich task before being fine-tuned on a downstream task that has not been seen during the pre-training. Pre-training is typically made for computer vision via supervised learning on large labeled training data such as ImageNet (Russakovsky et al. 2015). In natural language processing, pre-training is usually

done using unsupervised objective on large text corpora. Therefore, due to the availability of large scale unlabeled data on the Web, this approach has achieved state-of-the-art results in most NLP tasks such as question answering, text summarization and among others. Furthermore, transfer learning from pre-trained sentence embedding models has also demonstrated substantial gains in text representation learning. This section will briefly describe some popular pre-trained sentence embedding models. For the reader interested in a broader literature review for transfer learning techniques in NLP tasks, a recent survey by Ruder et al. (2019), is recommended.

Traditional sentence embedding models construct sentences embeddings by computing the weighted average of their words embeddings vectors. In this context, Ethayarajh (2018) has introduced the unsupervised smoothed inverse frequency model that uses an average of word embeddings and principal component removal to generate sentence representation. Iyyer et al. (2015) have introduced a Deep Average Network (DAN) that takes as input the average of words embeddings and bi-grams, which then passed through a feed forward neural network to produce the final sentence embedding vector. DAN has achieved good performances on many NLP tasks including text classification (Cer et al. 2018). Furthermore, pre-trained models based on recurrent neural networks, long short-term memory and gated recurrent neural networks in particular have also shown impressive results (Conneau et al. 2017; Howard and Ruder 2018).

Nevertheless, a recent trend is to use models based on the "Transformer" architecture (Vaswani et al. 2017). The Transformer is considered the state-of-the-art architecture for language understanding; it mainly based on self-attention instead of recurrent layers in an encoder-decoder model. Sentence embedding models that are based on the transformer architecture have achieved state-of-the-art results in text representation learning. For instance, Cer et al. (2018)

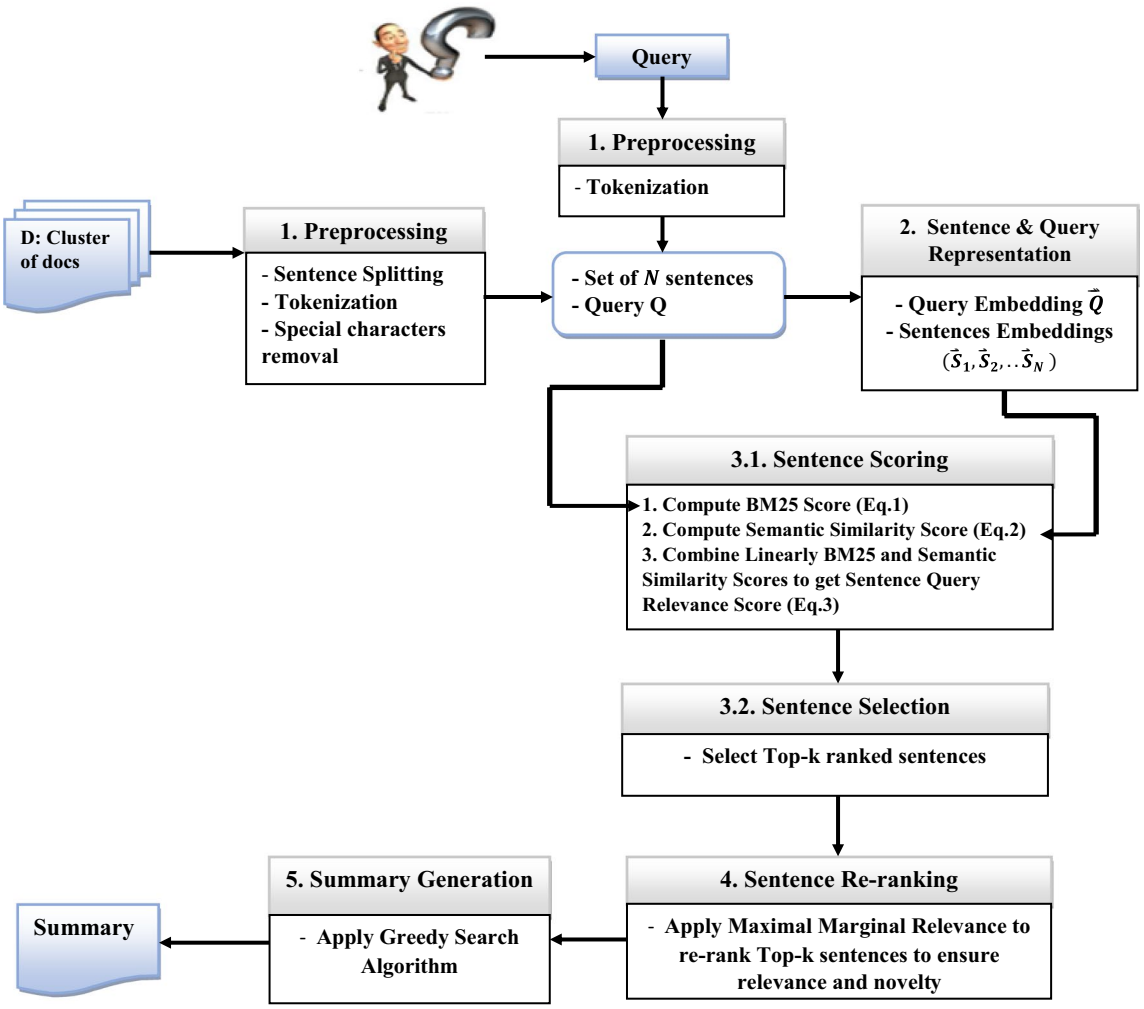

Fig. 1 Overall architecture of the proposed query-focused multi-document summarization system (QF-MDS)

have presented a universal sentence encoder that builds sentence embedding representation using the encoding sub-graph of the transformer architecture proposed by Vaswani et al. (2017). This sub-graph computes context-aware representations of words in a sentence by considering both the ordering and the identity of all other words. Furthermore, Devlin et al. (2019) have developed the Bidirectional Encoder Representations from Transformers (BERT) model based on a multi-layer bidirectional transformer with attention mechanisms. Raffel et al. (2019) have developed T5 (Text-to-Text-Transfer-Transformer) model based on the transformer with attention mechanisms. T5 is an encoder-decoder model pre-trained based on a multi-task learning paradigm. It uses a mixture of supervised and unsupervised tasks where each task is converted into text-to-text format. More recently, Brown et al. (2020) have presented the latest breakthrough language model GPT-3 (Generative Pre-trained Transformer 3), which is the third version of OpenAI GPT (Radford et al. 2018). GPT-3 is a powerful autoregressive language model based on the transformer architecture where the number of parameters has swelled to 175 billion.

Pre-trained sentence embedding models have been widely used in natural language processing tasks including generic text summarization. For instance, Joshi et al. (2019) have proposed an extractive unsupervised method for generic single text summarization based on deep auto-encoders and Skip-Thoughts sentence embedding model (Kiros et al. 2015) for sentence representation learning. Liu and Lapata (2019) have introduced BertSum, a novel document-level encoder for both extractive and abstractive generic text summarization based on the pre-trained BERT model (Devlin et al. 2019). Lewis et al. (2020) have introduced BART model that jointly pre-trains a seq2seq model by combining a bidirectional encoder and an auto-regressive decoder. It has been fine-tuned on an abstractive generic text summarization dataset and achieved the current state-of-the-art result in terms of ROUGE scores. Zhong et al. (2020) have developed MATCHSUM, a summary-level framework that formulates generic extractive summarization task as a semantic text matching problem using the siamese BERT model (Devlin et al. 2019). Xu and Lapata (2020) have proposed a coarse-to-fine modeling framework for extractive multi-document summarization that uses BERT model (Devlin et al. 2019) to capture the semantic relations between queries and document sentences. Recently, Lamsiyah et al. (2020) have introduced an unsupervised method for generic multi-document summarization based on sentence embedding models and centroid approach; they have provided an empirical analysis of several sentence embedding models and shown that the use of these methods improves the performances of extractive unsupervised summarization task.

In contrast to the previous methods, we propose a simple and effective unsupervised method that leverages transfer learning from pre-trained sentence embedding models to improve the performance of query-focused multi-document summarization task. The idea is justified by the fact that transfer learning helps in saving time and computational power as well as benefiting from knowledge learned across other natural language understanding tasks.

3 Proposed method

In this section, we first define the problem of query-focused multi-document summarization task. Then, we depict in details the main steps of the proposed solution.

Given a cluster of documents D consisting of a set of N sentences, a user query represented by a sentence Q, and a constraint on the summary length L. The aim of the proposed extractive query-focused multi-document summarization system is to find a query-relevant and non redundant summary *Summary* for the cluster D, such that the *Summary* $\subseteq N$ and the summary length L is reached.

Specifically, the process of our method is divided into five main steps: (1) preprocessing, (2) sentence and query representation, (3) sentence scoring and selection based on query relevance, (4) sentence re-ranking for redundancy removal, and (5) summary generation. The proposed method's overall architecture is illustrated in Fig. 1 while its procedure is described in Algorithm 1.

Algorithm 1 Pseudo code of the proposed method

Input: Q: an input query; $Sents$: an array of N sentences of the cluster D; $embeddingModel$: sentence embedding model; L: summary length; τ: sentence similarity threshold, α and λ: weighting parameters
Output: *summary*
Begin
  $queryRelevanceScores$: N-dimensional arrays ▷ Local variable
  $sentVectors \leftarrow embeddingModel.embedSentences(Sents)$ ▷ Sentences embeddings
  $queryVector \leftarrow embeddingModel.embedQuery(Q)$ ▷ Query embedding
  for $i \leftarrow 1$ to N **do**
    $BM25Scores[i] \leftarrow RSV_{BM25}(Sents[i], Q)$
    $semanticSimilarityScores[i] \leftarrow RSV_{Sim}(sentVectors[i], queryVector)$
    $queryRelevanceScores[i] \leftarrow \alpha * BM25Scores[i] + (1 - \alpha) * semanticSimilarityScores[i]$
  end for
  $top_k, indices \leftarrow sentencesSelection(Sents, queryRelevanceScores, k)$ ▷ select top k ranked sentences
  $sentVectors \leftarrow sentVectors[indices]$
  $summary \leftarrow []$
  $length \leftarrow 0$
  while $length < L$ **do** ▷ summary generation
    $MMR \leftarrow []$
    for $i \leftarrow 1$ to Length(top_k) **do**
      $MMR[i] \leftarrow score_{MMR}(sentVectors[i], queryVector, summary, \lambda)$
    end for
    $candidateSentence \leftarrow max(MMR)$
    $top_k \leftarrow top_k - candidateSentence$
    if $RSV_{Sim}(candidateSentence, sumary) < \tau$ **then** ▷ greedy selection
      $summary \leftarrow summary + candidateSentence$
      $length \leftarrow length + Length(candidateSentence)$
    end if
  end while
  return *summary*
End

3.1 Preprocessing

Formally, given a cluster D containing n documents $D = [d_1, d_2, \ldots, d_n]$, we split each document d_i in the cluster D into a set of sentences using the open-source software library for Advanced Natural Language Processing spaCy.[1] Then, we use the Natural Language Toolkit NLTK[2] and regular expressions to convert all the words of the obtained sentences to lower case and to remove special characters, XML/HTML tags, and redundant whitespace. Moreover, it is worth mentioning that the user query Q is also a sentence. Finally, we obtain a cluster D of N sentences, denoted as $D = [S_1, S_2, \ldots, S_N]$, and an associated query Q.

3.2 Sentence and query representation

Sentence and query representation is considered as a fundamental cornerstone task in query-focused multi-document summarization. As mentioned in Sect. 1, bag-of-words and word embedding representations are not able to carry the meaning and the semantic of a full sentence in one vector because they do not consider both the ordering of words as well as the semantic and syntactic relationships between them. Therefore, in our method, we exploit the potential of universal pre-trained sentence embedding models to encode clusters' sentences and users' queries into fixed-length vectors assuming that sentences with similar meanings have similar vectors, and simultaneously, sentences with different meanings have different vectors.

We consider three sentence embedding models including the non-parameterized unsupervised smoothed inverse frequency (uSIF) (Ethayarajh 2018), which does not need any external data and any training only pre-trained word embedding vectors. The two parameterized sentence encoders DAN and Transformer (Cer et al. 2018) require training to optimize their parameters. uSIF model utilizes a pre-trained word vector model, tuned on the ParaNMT-50 dataset (Wieting and Gimpel 2018), to generate word embedding vectors. Then, it creates sentence embedding vectors using the weighted average of word embedding vectors followed by a modification with singular vector decomposition and an unsupervised random walk algorithm. The universal sentence encoder (USE-DAN) is trained with a deep averaging network (DAN) (Iyyer et al. 2015) where word and bi-grams embedding vectors are averaged together and then passed through a deep neural network for sentence representation

[1] https://spacy.io/.
[2] https://www.nltk.org/.

learning. The universal sentence encoder Transformer (USE-Transformer) builds sentences embeddings using the encoding sub-graph of the transformer model (Vaswani et al. 2017). This encoding sub-graph computes context aware representations of words in a sentence based on the attention mechanism, which allows to take into account both the ordering and the identity of all the other words. Then, it converts the context aware word representations to a fixed length sentence encoding vector by computing the element-wise sum of the representations at each word position.

The universal sentence encoders USE-DAN and USE-Transformer have been trained on unlabeled data selected from Wikipedia, web news, web question-answer pages and discussion forums, and then fine-tuned on the human labeled SNLI dataset (Bowman et al. 2015), essentially for learning the semantic similarity between a pair of sentences. It's worth noticing that the USE-Transformer model has been further fine-tuned on the SQuAD question answering dataset (Rajpurkar et al. 2016). Hence, we benefit from transfer learning abilities by leveraging the knowledge learned from these supervised natural language understanding tasks (SNLI, SQuAD, and ParaNMT-50M) to improve the performance of the unsupervised QF-MDS task.

Formally, given a cluster of documents D consisting of N sentences, denoted as $D = [S_1, S_2, \ldots, S_N]$, and the input query Q. We use the three sentence embedding models as features extractors to generate embedding vectors for the input query Q and for each sentence S_i in the cluster D, which are denoted as $\overrightarrow{Q}$ and $\overrightarrow{S_i^D}$ respectively.

3.3 Sentence scoring and selection based on query relevance

In this step, we assign a score for each sentence S_i in the cluster D based on its relevance to the query Q using two different metrics the $BM25$ model and the semantic similarity. Then, we select the $top-k$ ranked sentences according to their final score obtained by linearly combining the two latter metrics. Sentence scoring as well as sentence selection methods are sequentially described in the following subsections:

3.3.1 Sentence scoring

Let $D = [S_1, S_2, \ldots, S_N]$ denote a cluster of documents containing N sentences, and Q the input query. We use the $BM25$ model and the semantic similarity to measure each sentence's relevance S_i in the cluster D to the input query Q.

BM25 Model (Robertson et al. 1995) is considered as one of the most popular probabilistic information retrieval models. It is based on a binary independence assumption where a query term's weight is computed using both its within-sentence term frequency and query term frequency. The relevance score for a sentence S_i given a query Q is defined by:

$$RSV_{BM25}(S_i, Q) = \sum_{w \in S_i \cap Q} \frac{(k_1+1).x_w^{S_i}}{K + x_w^{S_i}} * \frac{(k_3+1).x_w^{Q}}{k_3 + x_w^{Q}} * \log \frac{N - N_w + 0.5}{N_w + 0.5} \quad (1)$$

where $K = k_1.((1-b) + b.\frac{l_{S_i}}{l_{avg}})$ is the parameter for the within sentence frequency normalization, k_1 is a positive tuning parameter that calibrates the sentence term frequency scaling, b is the parameter for normalizing the sentence length, and k_3 is the parameter for weighting the query term frequency.

Semantic similarity metric is used to deal with term mismatch problem, occurred when relying on exact term matching between the query and the cluster's sentences. It measures the degree to which a sentence S_i and a query Q carry the same meaning by computing the cosine similarity between their embedding vectors. The relevance score of a sentence S_i to the query Q using the semantic similarity metric, is formally defined as follows:

$$RSV_{Sim}(S_i, Q) = \frac{\overrightarrow{S_i^D} \cdot \overrightarrow{Q}}{||\overrightarrow{S_i^D}|| \cdot ||\overrightarrow{Q}||} \quad (2)$$

Where $\overrightarrow{S_i^D}$ is the embedding vector of a sentence S_i in the cluster D, and $\overrightarrow{Q}$ is the embedding vector of the input query Q.

3.3.2 Sentence selection

In this step, we apply a retrieval method to select the top-k relevant sentences to the input query Q. Given the RSV_{BM25} and RSV_{Sim} scores of each sentence S_i in the cluster D, we assume that relevant sentences to the query Q are those that maximize the weighted sum of these two scores. As defined in Equation 3, we combine linearly the RSV_{BM25} and RSV_{Sim} scores to get the final query relevance score of a sentence S_i, denoted as $score_{Relevance}(S_i)$. Then, based on the obtained scores, we iteratively select the top-k ranked sentences such as $k \in \{50, 100\}$. The top-k selected sentences, denoted as top-$k = \{S_1, S_2, \ldots, S_k\}$, are considered as a set of candidate sentences for the final summary.

$$score_{Relevance}(S_i) = \alpha * RSV_{BM25}(S_i, Q) + (1-\alpha) * RSV_{Sim}(S_i, Q) \quad (3)$$

Where $\alpha \in [0, 1]$ with constant steps of 0.1.

3.4 Sentence re-ranking for redundancy removal

Given the top-$k = \{S_1, S_2, \ldots, S_k\}$ selected sentences, we re-rank these sentences in order to produce a summary that combines two main factors the *query relevance* and

Table 2 Statistics of DUC'2005–2007 Datasets

Datasets	Clusters	Num docs	Sentences	Queries	Summary length	Num gold summaries	Data Source
DUC'2005	50	32	45,931	50	250	4	TREC
DUC'2006	50	25	34,560	50	250	4	AQUAINT
DUC'2007	45	25	24,282	45	250	4–9	AQUAINT

Num docs indicates the average number of documents in each cluster. *Summary length* corresponds to the number of words in gold summaries. *Num gold summaries* corresponds to the number of human summaries provided for each cluster

the *novelty*. The former measures how relevant a sentence is to the given query while the latter allows dealing with redundancy and producing summaries with good information diversity. To this end, we use the Maximal Marginal Relevance method (Carbonell and Goldstein 1998), formally defined in Eq. 4. Given a sentence S_i in top-$k = \{S_1, S_2, \ldots, S_k\}$, we first compute the relevance of S_i according to the query Q using the cosine similarity between their corresponding embedding vectors $\overrightarrow{S_i}$ and $\overrightarrow{Q}$, and then we calculate its cosine similarity with the already selected sentences as summary. Finally, we combine linearly these two scores (relevance, novelty) to get the MMR score of the sentence S_i, denoted as $score_{MMR}(S_i)$. Where S_i has a high marginal relevance if it is relevant to the query and contains minimal similarity to previously selected sentences.

$$\begin{aligned} score_{MMR}(S_i) \; = \mathrm{Argmax}_{S_i \in \text{top-}k \setminus Sum} \Big[& \lambda RSV_{Sim}(S_i, Q) \\ & - (1-\lambda) \max_{S_j \in Sum} CosSim(\overrightarrow{S_i}, \overrightarrow{S_j}) \Big] \\ & 1 \le i \le k, i \ne j \end{aligned} \tag{4}$$

Where, $RSV_{Sim}(S_i, Q_D)$ is the relevance score of S_i according to the query Q (Equation 2), $CosSim(\overrightarrow{S_i}, \overrightarrow{S_j})$ is the cosine similarity between the embedding vectors of the current sentence S_i and the already selected sentences as summary sentences S_j, top-k is a set of sentences selected in the previous step, *Sum* subset of sentences in top-k already selected as summary sentences, top-$k \setminus Sum$ set of unselected sentences in top-k, and λ is an interpolation coefficient in range [0, 1] with constant steps of 0.1.

3.5 Summary generation

Finally, after re-ranking the $top - k$ sentences, we apply a greedy search algorithm to produce the final summary with respect to the summary length L (for DUC'2005-2007 datasets, $L = 250$ words). We add a new sentence to the current summary if the length limit L is not reached and the similarity between the current sentence and the already selected summary sentences is below a threshold τ.

4 Experimental results

In this section, we first present a brief description of the used datasets, evaluation metrics, and the experimental setup. Then, we provide a comparative analysis of the obtained results, intending to verify the following hypotheses:

- *Hypothesis H1* The proposed method is effective for query-focused multi-document summarization as compared with other recent state-of-the-art extractive QF-MDS methods.
- *Hypothesis H2* The use of sentence embedding models has shown to be effective for unsupervised query-focused multi-document summarization.
- *Hypothesis H3* The combination of *BM25* and *semantic similarity* metrics improves sentence scoring method and leads to significant improvements.

4.1 Experimental datasets and evaluation metrics

We evaluate the proposed method on three standard benchmarks for query-focused multi-document summarization, namely DUC'2005-2007 datasets. DUC (Document Understanding Conference) datasets are created by NIST[3] (National Institute of Standards and Technology) and considered as the widely used corpora for evaluating text summarization. Each dataset consists of a set of clusters, each cluster contains an average of 25 English news articles and accompanied with a query. The query is represented by a sentence Q, which consists of the main topic followed by additional questions indicating the aspects that should the summary cover; e.g.:

> *"New Hydroelectric Projects. What hydroelectric projects are planned or in progress and what problems are associated with them? "*

In DUC'2006–2007 datasets, each cluster has 4 human-written summaries provided by different experts, while in DUC'2005 is about 4–9 summaries. Note that each summary's length is limited to 250 words, as required in DUC evaluations. Table 2 summarizes some basic statistics of DUC'2005-2007 datasets.

[3] https://duc.nist.gov/

In our experiments, we produce a summary for each cluster in DUC'2005-2007 datasets, which is then compared with the reference summaries using the popular ROUGE method (Lin 2004). Specifically, ROUGE-N (ROUGE-1 and ROUGE-2) and ROUGE-SU4. As defined in Eq. 5, ROUGE-N measures the similarity between the systems summaries and a collection of summaries models (human summaries) based on the N-gram overlap. While ROUGE-SU4 measures skip-bigram overlap between a system summary and a set of reference summaries with a max distance of four words.

$$ROUGE - N = \frac{\sum_{S \in (RS)} \sum_{N-gram \in (S)} match(N - gram)}{\sum_{S \in (RS)} \sum_{N-gram \in (S)} Count(N - gram)} \quad (5)$$

where N stands for the length of $N - gram$, $Count(N - gram)$ is the number of $N - gram$ in the Reference Summary RS, and $match(N - grams)$ is the maximum number of $N - gram$ that occur in both reference summary RS and the candidate summary S. We report the recall of ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-SU4 (R-SU4) using ROUGE toolkit (version 1.5.5). We adopt the same ROUGE settings[4] that are used on DUC'2005-2007 datasets for evaluating extractive query-focused multi-document summarization systems.

4.2 Experimental setup

We consider three sentence embedding models including uSIF, USE-DAN, and USE-Transformer. Each model is designed to encode the input sentences and queries into dense vectors with a specific dimension. For instance, uSIF model generates embedding vectors of 300 dimensions, while USE-DAN and USE-Transformer models provide embedding vectors of 512 dimensions. Moreover, we adopt the feature-based approach where the sentence embedding model is applied on the cluster of the documents to be summarized without any additional fine-tuning; in other words, we extract sentences and queries embedding vectors and use them directly in our method. The proposed method has been developed using Python, relying on TrecTools[5] library and the available implementation of uSIF[6] model as well as the pretrained USE-DAN[7] and USE-Transformer.[8] models

We have used four hyperparameters to get the final summary, namely α, λ, τ, and the number of top-k selected sentences k. Where, α is comprised between [0, 1] with constant steps of 0.1, λ and the threshold τ are comprised between [0.5, 0.95] with constant steps of 0.05, and $k \in \{50, 100\}$. To determine the values of these hyperparameters, we built a small held-out set by shuffling and randomly sampling 20 clusters from the DUC'2006 dataset. Then, we performed a grid search on the held-out set that gave us a total of 2000 feasible combinations. Accordingly, the obtained values of the hyperparameters are 0.3, 0.9, 0.85, and 50 for α, λ, τ, and k, respectively. Finally, we generated the summaries of DUC'2005 and DUC'2007 datasets and the other 30 clusters of DUC'2006 dataset.

4.3 Comparison with state-of-the-art methods

In order to examine the effectiveness of the proposed method (*Hypothesis H1*), we compare its performance on the standard DUC'2005-2007 datasets with the existing state-of-the-art methods for query-focused multi-document summarization including recent supervised deep learning based methods. Table 3 displays ROUGE recall scores of our method (denoted as **uSIF-Sum**, **USE-DAN-Sum**, or **USE-Transformer-Sum** according to the sentence embedding model that is used) and those of the other state-of-the-art methods on the three DUC'2005–2007 datasets. Note that, for the state-of-the-art methods, we present the reported results in their corresponding papers.

The first set of analysis is conducted to compare our method with different extractive unsupervised QF-MDS methods, described as follows: **BI_PLSA** (Shen et al. 2011) system applies a probabilistic latent semantic analysis method to simultaneously cluster and summarize documents. **HierSum** (Haghighi and Vanderwende 2009) is based on a hierarchical LDA-style for representing content specificity as a hierarchy of topic distributions. **SpOpt** (Yao et al. 2015) produces summaries using sparse optimization with a decomposable convex objective function. **QODE** (Zhong et al. 2015) is a deep learning based system that combines the restricted bolzmann machines and dynamic programming. **DPRQSum** (Wu et al. 2019) system is based on dual pattern-enhanced models for representing topical significance and query relevance for documents and sentences. **wAASum** (Canhasi and Kononenko 2014) is a graph-based system that uses a weighted archetypal analysis factorization method to estimate the importance of sentences in a cluster of documents. Finally, **CES** (Feigenblat et al. 2017) and **Dual-CES** (Roitman et al. 2020) systems are both based on the cross entropy method and considered as the best unsupervised query-focused multi-document summarization systems on DUC'2005-2007 datasets. The first block of Table 3 reports R-1, R-2, and R-SU4 scores of these methods.

As shown in Table 3, on DUC'2005–2006 datasets and based on R-1 and R-2 measures, the three variants of our method (uSIF-Sum, USE-DAN-Sum, and USE-Transformer-Sum) have outperformed PI-PLSA, HierSum, SpOpt, DPRQ-Sum, and QODE while achieving comparable performances to wAASum, CES and Dual-CES methods. On DUC'2007 dataset, our method has achieved comparable results to all the other methods except CES and Dual-CES, which have shown

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

[5] https://pypi.org/project/trectools/.

[6] https://github.com/kawine/usif.

[7] https://tfhub.dev/google/universal-sentence-encoder/4.

[8] https://tfhub.dev/google/universal-sentence-encoder-qa/3.

very high performances in terms of R-1 and R-2 scores. In terms of R-SU4 evaluation measure, our method based on the three used sentence embedding models has achieved the best performances; it has outperformed all the systems that we compared with on the three DUC'2005–2007 benchmarks. In detail, the USE-Transformer-Sum has obtained an improvement of 1.39%, 1.36%, and 0.63% with respect to Dual-CES system using DUC'2005-2007 datasets, respectively.

To further prove the efficiency of our method, we compare its performance with some of the state-of-the-art supervised methods, described as follows: **HybHSum** (Celikyilmaz and Hakkani 2010) formulates query-focused summarization as a prediction problem using a generative model for pattern discovery and a regression model for inference. **Ensemble-Sys-AOR** (Valizadeh and Brazdil 2015) combines an ensemble summarizing system and actor-object relationships between sentences to generate summaries. Finally, **AttSum** (Cao et al. 2016), **CRSum-SF** (Ren et al. 2017), and **SRSum** (Ren et al. 2018) methods use deep learning models with attention mechanisms. The second block of Table 3 reports R-1, R-2, and R-SU4 scores of these methods.

From Table 3, we can see that our method has outperformed AttSum method on DUC'2005 dataset and achieved comparable results with it on DUC'2006-2007 datasets. Moreover, in terms of R-1 and R-2 measures and using DUC'2005-2006 datasets, our method has shown comparable performances to strong methods including Ensemble-Sys-AOR, HybHSum, SRSum, and CRSum-SF. However, on DUC'2007 dataset, the latter methods achieved better performances than our method, especially the HybHSum model which yielded the best R-1 score. Finally and in terms of R-SU4 measure, our method's three variants have shown far better performances than all the other methods on the three datasets. For instance, USE-Transformer-Sum has obtained an increment of 2.18% and 1.34% with respect to the HybHSum system on DUC'2006-2007 datasets.

Therefore, the overall obtained results on the three datasets DUC'2005-2007 have demonstrated that the unsupervised Dual-CES system, the supervised SRSum system, and our method (uSIF-Sum, DAN-Sum, and Transformer-Sum) are the best performing systems in terms of R-1, R-2, and R-SU4 measures, respectively.

4.4 Effectiveness of sentence embedding representation

The second set of analysis's objective is to evaluate the use of sentence embedding representation on unsupervised query-focused multi-document summarization task (examine *Hypothesis H2*). To this end, we have implemented our method using three different representations: bag-of-words using *TF-IDF* weighting scheme (Ramos 2003), word embeddings using GloVe model (Pennington et al. 2014), and sentence embeddings using three different models (uSIF, USE-DAN, and USE-Transformer). ROUGE recall scores of these methods (denoted as **TF-IDF-Sum**, **GloVe-Sum**, **uSIF-Sum**, **USE-DAN-Sum**, or **USE-Transformer-Sum** according to the representation that is used) on DUC'2005–2007 datasets are summarized in Table 4. A

Table 3 ROUGE recall results of the proposed method and state-of-the-art systems on DUC'2005–2007 datasets

	DUC'2005			DUC'2006			DUC'2007		
	R-1	R-2	R-SU4	R-1	R-2	R-SU4	R-1	R-2	R-SU4
BI-PLSA	36.02	6.76	–	39.38	8.49	–	–	–	–
HierSum	–	–	–	40.1	8.6	14.3	42.4	11.8	16.7
SpOpt	–	–	–	39.96	8.68	14.22	42.36	11.1	16.47
QODE	37.51	7.75	13.88	40.15	9.28	14.79	42.95	11.63	16.85
wAASum	39.45	7.97	14.20	42.38	9.17	16.71	–	–	–
DPRQSum	–	–	–	40.55	9.22	14.96	43.40	11.68	17.02
CES	40.33	7.94	13.89	43	9.69	15.63	45.43	12.03	17.5
Dual-CES	**40.82**⋆	**8.07**	**14.13**	**43.94**⋆	**10.09**	**15.96**	**46.02**⋆	**12.53**	**17.91**
Ensemble-Sys-AOR	–	–	–	–	9.77	15.28	–	12.70	17.46
AttSum	37.01	6.99	–	40.9	9.4	–	43.92	11.55	–
HybHSum	–	–	–	43	9.1	15.1	**45.6**	11.4	17.2
CRSum-SF	39.52	8.41	–	41.7	10.03	–	44.6	12.48	–
SRSum	**39.83**	**8.57**⋆	–	**42.82**	**10.46**⋆	–	45.01	**12.8**⋆	–
uSIF-Sum	37.97	7.85	14.39	40.62	8.98	16.29	42.25	10.59	17.24
USE-DAN-Sum	38.94	7.89	15.07	41.74	9.26	16.56	43.04	10.74	18.15
USE-Transformer-Sum	**39.79**	**8.27**	**15.52**⋆	**42.8**	**9.39**	**17.28**⋆	**43.54**	**11.42**	**18.54**⋆

The results of the state-of-the-art systems are taken from their original articles. The symbol "–" means that the results are not reported in their respective works. The highest performance of R-1, R-2, and R-SU4 for each of the group of methods is printed in boldface. The best performing method for each measure is indicated by ⋆

concrete example of the generated summaries of these methods with the reference summaries is illustrated in Table 6 in Sect. 1. We also performed the paired student's t-test (Dietterich 1998) between the ROUGE scores of these methods and attached a superscript to the performance number in the table when the $p-$value < 0.05.

Table 4 shows clearly that on the three benchmarks, USE-DAN-Sum, and USE-Transformer-Sum have significantly outperformed *TF-IDF*-Sum and GloVe-Sum methods for all evaluation measures. For instance, on DUC'2005 corpus, the USE-Transformer-Sum achieves an improvement of 4.77%, 1.02%, and 2.36% with respect to *TF-IDF*-Sum in terms of R-1, R-2, and R-SU4, respectively. Moreover, uSIF-Sum model has also achieved promising results and lead to significant improvements over the *TF-IDF*-Sum and GloVe-Sum methods in terms of R-1 and R-SU4 measures. These noteworthy results verify that incorporating sentence embedding representation to the semantic similarity and the maximal marginal relevance functions can improve the performance of unsupervised query-focused multi-document summarization task.

Several experiments were conducted to compare the three different sentence embedding models (uSIF, USE-DAN, and USE-Transformer). As shown in Table 4, on the three DUC'2005–2007 benchmarks, the USE-Transformer-Sum model has achieved the best performances and significantly outperformed uSIF-Sum and USE-DAN-Sum models for most evaluation measures (R-1, R-2, and R-SU4). Moreover, in terms of R-1 and R-SU4 scores, the USE-DAN-Sum has yielded significant improvements over the uSIF-Sum model, while they have achieved comparable performances in terms of R-2 score. Therefore, the overall obtained results on the three DUC'2005–2007 datasets have demonstrated that the three sentence embedding models have shown to be effective for the query-focused summarization task. Indeed, both USE-DAN and USE-Transformer encoding models have been trained using a large amount of unsupervised data drawn from various web sources documents, and for boosting the performance, the unsupervised learning is augmented with supervised training on SNLI dataset (Bowman et al. 2015). The latter dataset is considered as one of the largest and high-quality labeled corpus, designed for textual entailment tasks. Furthermore, we observe that transfer learning from the USE-transformer based sentence encoder performs better than transfer learning from the USE-DAN encoder. This can be due to their different architectures as well as the used training datasets. The USE-Transformer is further fine-tuned on SQuAD dataset (Rajpurkar et al. 2016), created for the question-answering systems. Moreover, uSIF model has also shown promising results. Indeed, uSIF model has been tested on the SemEval Semantic Textual Similariy (STS) task and has achieved the state-of-the-art performance on this task (Ethayarajh 2018). This task's goal lies in computing the semantic similarity between a pair of sentences, applying the cosine similarity between their corresponding vectors. The obtained results using uSIF model contrast this finding, since in our method we also use semantic similarity in both sentences scoring and re-ranking.

4.5 Effectiveness of combining the *BM*25 and semantic similarity metrics

As an additional contribution of this work is the combination of *BM*25 and *semantic similarity* sentence scoring metrics, we conducted a set of experiments to examine hypothesis *H3*: (1) *Does the combination of these two metrics improve the performance of sentence scoring method?*

To address the latter empirical question, we have performed three runs on DUC'2005–2007 datasets, described as follows:

- **Run 1:** top-k sentences are selected based on the *BM*25 score RSV_{BM25};
- **Run 2:** top-k sentences are selected based on the *semantic similarity* score RSV_{Sim};
- **Run 3:** top-k sentences are selected using the linear combination of both the RSV_{BM25} and RSV_{Sim} scores.

Table 5 summarizes the obtained ROUGE recall scores (R-1, R-2, and R-SU4) of the three runs on DUC'2005–2007 datasets.

Table 4 Comparison results of *TF-IDF*-Sum, GloVe-Sum, uSIF-Sum, USE-DAN-Sum, and USE-Transformer-Sum methods on the query-focused multi-document summarization datasets DUC'2005–2007

	DUC'2005			DUC'2006			DUC'2007		
	R-1	R-2	R-SU4	R-1	R-2	R-SU4	R-1	R-2	R-SU4
TF-IDF-Sum[1]	35.02	7.25	13.16	35.79	7.49	13.38	36.32	10.22	13.88
GloVe-Sum[2]	37.66^{1}	7.67	14.05^{1}	39.50^{1}	7.88	15.03^{1}	41.92^{1}	10.32	16.33^{1}
uSIF-Sum[3]	37.96^{1}	7.85^{1}	$14.39^{1\text{-}2}$	$40.62^{1\text{-}2}$	$8.98^{1\text{-}2}$	$16.29^{1\text{-}2}$	$42.25^{1\text{-}2}$	10.59	$17.24^{1\text{-}2}$
USE-DAN-Sum[4]	$38.94^{1\text{-}3}$	7.89^{1}	$15.07^{1\text{-}3}$	$41.74^{1\text{-}3}$	$9.26^{1\text{-}2}$	$16.56^{1\text{-}2}$	$43.04^{1\text{-}3}$	$10.74^{1\text{-}2}$	$18.15^{1\text{-}3}$
USE-Transformer-Sum[5]	$39.79^{1\text{-}4}$	$8.27^{1\text{-}4}$	$15.52^{1\text{-}4}$	$42.8^{1\text{-}4}$	$9.36^{1\text{-}3}$	$17.28^{1\text{-}4}$	$43.54^{1\text{-}4}$	$11.42^{1\text{-}4}$	$18.54^{1\text{-}4}$

For denoting statistical significance results, the superscripts *number* indicates significant improvement ($p-$value < 0.05) over the method that has the same superscript *number* attached. The interval $i-j$ indicates a significant improvement over models that have a superscript *number* attached ranging from i to j

From Table 5, it seems clear that the *BM*25 metric (Run 1) has achieved modest results in comparison with the *semantic similarity* metric (Run 2). This can be explained by the fact that *BM*25 model suffers from the problem of term mismatch where query terms may not occur in relevant sentences. On the three DUC'2005–2007 benchmarks, the *semantic similarity* in conjunction with the three sentence embedding models (uSIF, USE-DAN, and USE-Transformer) has shown promising results for the most evaluation measures (R-1, R-2, and R-SU4) and lead to significant improvements over *BM*25 model. The latter demonstrates sentence embedding models' effectiveness in capturing sentences' semantic meaning. Moreover, the combination of the *semantic similarity* and the *BM*25 (Run 3) has yielded significant improvements over Run 1 and Run 2 for the most used evaluation measures in all benchmarks. For instance, for DUC'2007 dataset, Run 3 using uSIF model obtains an increment of 0.71%, 0.51%, and 0.19% with respect to Run 2 for R-1, R-2, and R-SU4 metrics respectively. Additionally, for R-1 measure, Run 3 achieves an increment of 0.5% and 0.17% with respect to Run 2 using USE-DAN and USE-Transformer models respectively. *BM*25 model has contributed significantly to improving the performance of sentence scoring function. The comparison results show that the *BM*25 and *semantic similarity* metrics are complementary to each other.

5 Conclusion

In this paper, we proposed an unsupervised method for query-focused multi-document summarization based on transfer learning from pre-trained sentence embedding models. First, sentence embedding models are exploited to represent the cluster's sentences and the users 'queries into fixed dense vectors, which are used to compute the semantic similarity between the cluster's sentences and the user's query. Then, the semantic similarity and the BM25 model are linearly combined to select the top-k ranked sentences based on their relevance to the input query. Finally, the maximal marginal relevance criterion is used to re-rank the top-k selected sentences to produce a query-relevant summary that maximizes relevant information and minimizes redundancy.

We performed an extensive experimental analysis to validate the robustness of the proposed query-focused multi-document summarization method. In particular, several experiments were conducted on DUC'2005–2007 datasets to evaluate each sentence embedding model exploited in our method and assess the impact of combining the semantic similarity function and the BM25 model. The experimental results showed that the use of sentence embedding representation and the combination of the semantic similarity and the BM25 model, have considerably improved the results for all evaluation measures (R-1, R-2, and R-SU4) in all benchmarks. We compared our method with several state-of-the-art query-focused multi- document summarization systems including recent supervised deep learning based systems. The obtained results show that our method has achieved promising results. In particular, it has outperformed several systems and achieved comparable results to the best performing unsupervised systems (CES and Dual-CES) and even outperformed them in terms of R-SU4 measure. Besides, its summarization quality can reach that of strong supervised systems (AttSum, CRSum-SF, and SRSum).

Transfer learning from pre-trained sentence embedding models has shown to be effective for boosting the performance of the query-focused multi-document summarization. Hence, in the future work, we plan to investigate the potential of the newest models T5 and GPT-3 for text summarization task. Furthermore, we also plan to explore transfer learning abilities from pre-trained models for summary generation.

Example of generated summaries

Table 6 presents the generated summaries using TF-IDF-Sum, GloVe-Sum, uSIF-Sum, USE-DAN-Sum, and USE-Transformer-Sum for the query *D*307 of the DUC'2005 dataset.

Table 5 The obtained results of the sentence scoring metrics (*BM25* and *semantic similarity*) and their combination using three different sentence embedding models (uSIF, USE-DAN, and USE-Transformer) on the query-focused multi-document summarization datasets DUC'2005–2007

		uSIF			USE-DAN			USE-Transformer		
		R-1	R-2	R-SU4	R-1	R-2	R-SU4	R-1	R-2	R-SU4
DUC'2005	RUN 1	36.23	6.89	13.45	37.16	7.11	13.88	38.59	7.19	14.81
	RUN 2	37.81^{1}	7.68^{1}	14.31^{1}	38.55^{1}	7.62^{1}	14.76^{1}	39.65^{1}	8.21^{1}	15.5^{1}
	RUN 3	37.96^{1}	7.85^{1}	14.39	$38.94^{1\text{-}2}$	$7.89^{1\text{-}2}$	$15.07^{1\text{-}2}$	39.79^{1}	8.27^{1}	15.52^{1}
DUC'2006	RUN 1	39.01	7.44	15.02	39.69	8.14	15.24	40.29	8.40	15.72
	RUN 2	40.47^{1}	8.65^{1}	15.94^{1}	40.94^{1}	9.07^{1}	16.44^{1}	42.10^{1}	9.24^{1}	16.91^{1}
	RUN 3	40.62^{1}	$8.98^{1\text{-}2}$	$16.29^{1\text{-}2}$	$41.74^{1\text{-}2}$	$9.26^{1\text{-}2}$	16.56^{1}	$42.8^{1\text{-}2}$	$9.36^{1\text{-}2}$	$17.28^{1\text{-}2}$
DUC'2007	RUN 1	40.38	9.04	16.02	40.71	9.67	16.42	41.43	9.87	16.73
	RUN 2	41.54^{1}	10.08^{1}	17.05^{1}	42.54^{1}	10.41^{1}	17.84^{1}	43.37^{1}	11.10^{1}	18.11^{1}
	RUN 3	$42.25^{1\text{-}2}$	$10.59^{1\text{-}2}$	$17.24^{1\text{-}2}$	$43.04^{1\text{-}2}$	$10.74^{1\text{-}2}$	$18.15^{1\text{-}2}$	$43.54^{1\text{-}2}$	$11.42^{1\text{-}2}$	$18.54^{1\text{-}2}$

The subscripts 1, 2, and 3 indicate significant improvement ($p-$ value < 0.05) of Run 1, Run 2, and Run 3, respectively

Table 6 Example of generated summaries of TF-IDF-Sum, GloVe-Sum, uSIF-Sum, USE-DAN-Sum, and USE-Transformer-Sum. The corresponding reference summaries are presented in the next page of the appendix

Query

New Hydroelectric Projects. What hydroelectric projects are planned or in progress and what problems are associated with them?

TF-IDF-Sum

It is the most serious of all problems associated with the dam, he says. • Hydro-Quebec has always insisted that none of its new projects are driven entirely by export contracts

Nearly MDollars 6 bn has been raised on the domestic market this year for projects associated with Malaysia's plans to privatise a large part of its electricity industry

Knight Piesold says the scheme involves the multi-purpose development of the Ewaso Ngiro river and includes three separate hydroelectric projects, a river transfer scheme and an irrigation project

Railway electrification is planned.

The new agreement requires majority approval, at council level, only for projects that divert water out of the Mekong basin during the dry season

The country cannot neglect the value of its 42,000 MW of potential hydroelectric generating capacity. Sites have been chosen for the new cities and an overall broad plan mapped

However, large hydroelectric schemes cost around Dollars 1500 per kilowatt

While Great Whale will no longer be needed to supply power to New York in 1998 as originally planned, the utility still plans to press ahead with the CDollars 12 bn project once a thorough environmental assessment by the federal and provincial governments is complete

THE Mexican government has dropped plans to build a hydroelectric dam on the Rio Usumacinta, near the Guatemalan border

The bank is promoting the project at a time of growing international concern about evidence that big dams in developing countries often do not deliver expected economic benefits and sometimes cause unexpected problems

India was asked by the bank to improve detailed plans for resettling displaced villagers and to prepare a full study of the project's environmental effects.

GloVe-Sum

It is the most serious of all problems associated with the dam, he says

Hydro-Quebec has always insisted that none of its new projects are driven entirely by export contracts

Nearly MDollars 6 bn has been raised on the domestic market this year for projects associated with Malaysia's plans to privatise a large part of its electricity industry

Knight Piesold says the scheme involves the multi-purpose development of the Ewaso Ngiro river and includes three separate hydroelectric projects, a river transfer scheme and an irrigation project

The new agreement requires majority approval, at council level, only for projects that divert water out of the Mekong basin during the dry season

Railway electrification is planned

The fight centres on Kelantan's insistence that it is entitled to substantial compensation and revenues associated with the dam project, which will use water from the state's Pergau river

This is the first dam in a proposed series of six, and the largest, most controversial hydroelectric project to receive IFC financing

The country cannot neglect the value of its 42,000 MW of potential hydroelectric generating capacity

Sites have been chosen for the new cities and an overall broad plan mapped. Yacyreta, being built by Impregilio of Italy and Dumez of France, is the second big hydro-electric dam to be built on the Parana, and two more are planned

However, large hydroelectric schemes cost around Dollars 1500 per kilowatt

While Great Whale will no longer be needed to supply power to New York in 1998 as originally planned, the utility still plans to press ahead with the CDollars 12 bn project once a thorough environmental assessment by the federal and provincial governments is complete

uSIF-Sum

Knight Piesold says the scheme involves the multi-purpose development of the Ewaso Ngiro river and includes three separate hydroelectric projects, a river transfer scheme and an irrigation project

By contrast the next 5 years will see the emergence of a mixed bag of thermal, combined cycle, and hydroelectric projects, the latter comprising four major projects on the Karun river in the south-west, and a 1000 MW pumped storage plant at Siabisheh on the Caspian sea in the north

The Xiaolangdi water control project contracts for the main body, the flood-discharge structure and hydroelectric structure were awarded to a Yellow River joint venture

The World Bank's ability to continue backing large-scale power projects in the developing world is facing a crucial test as a result of environmentalists' opposition to a proposed Arun III hydroelectric scheme for Nepal

The consortium led by Philip Holzmann of Germany has been quick to point out that there are no plans for irrigation, only hydro-electric generation

Nearly MDollars 6 bn has been raised on the domestic market this year for projects associated with Malaysia's plans to privatise a large part of its electricity industry

In the 1940s US engineers, fresh from numerous 'big dam' projects in north America, first laid plans for a cascade of 'multi-purpose' dams along the river for electricity, flood control, irrigation and improved river transport

In an effort to spread its risks, Hydro-Quebec has begun encouraging co-generation projects, mostly in partnership with local pulp and paper mills

This is the first dam in a proposed series of six, and the largest, most controversial hydroelectric project to receive IFC financing.

Table 6 (continued)

USE-DAN-Sum

By next century, when the irrigation infrastructure and the 21 dams and 19 power plants are in place, the project will have cost a heady Dollars 32 bn

Since then the project has been endlessly debated and researched by politicians, engineers and hydro-electric specialists

By contrast the next 5 years will see the emergence of a mixed bag of thermal, combined cycle, and hydroelectric projects, the latter comprising four major projects on the Karun river in the south-west, and a 1000 MW pumped storage plant at Siabisheh on the Caspian sea in the north

Lengthy environmental reviews may delay two other northern Quebec projects, known as Eastmain-1 and Laforge-2

Knight Piesold says the scheme involves the multi-purpose development of the Ewaso Ngiro river and includes three separate hydroelectric projects, a river transfer scheme and an irrigation project. • The town will be re-built back from the river with some of the 18.5 bn yuan allocated to resettlement from the project's budget of 57 bn yuan

These include village-level micro-dams, medium-sized dams for towns and at least one large project—a Dollars 300 m, 140 MW project on the river Kaligandaki in central Nepal

The fall-out from Narmada and the criticism of Arun are causing worry within the Bank about its future involvement in large-scale energy projects

In the 1940s US engineers, fresh from numerous 'big dam' projects in north America, first laid plans for a cascade of 'multi-purpose' dams along the river for electricity, flood control, irrigation and improved river transport

Averaging the expense of a mix of thermal, gas-fired and hydroelectric plants at a conservative Dollars 1000 per installed kilowatt produces a future gross figure of about Dollars 3 bn a year of which local contracting and supply might account for 30–40 per cent

USE-Transformer-Sum

Because other privatised energy companies will be seeking funds, IFC's handling of Pangue is likely to set environmental standards for projects to follow

Knight Piesold says the scheme involves the multi-purpose development of the Ewaso Ngiro river and includes three separate hydroelectric projects, a river transfer scheme and an irrigation project

The south-east Anatolian project or Gap, as it sometimes better known, is currently the largest development project in the Mediterranean

The 600 km reservoir will be a tranquil body of water, twice the width of the present turbulent river, improved navigation being one of the project's proclaimed benefits

The funding for a controversial hydroelectric dam to be built in the heart of Malaysia's tropical rainforest will be generated from domestic sources, according to the company in charge of the project

Hydro-Quebec has always insisted that none of its new projects are driven entirely by export contracts

Nearly MDollars 6 bn has been raised on the domestic market this year for projects associated with Malaysia's plans to privatise a large part of its electricity industry

The fight centres on Kelantan's insistence that it is entitled to substantial compensation and revenues associated with the dam project, which will use water from the state's Pergau river

Lengthy environmental reviews may delay two other northern Quebec projects, known as Eastmain-1 and Laforge-2

The project will generate 84 bn kw/h of hydro-electric power a year, one-eighth of China's 1991 output

By contrast the next 5 years will see the emergence of a mixed bag of thermal, combined cycle, and hydroelectric projects, the latter comprising four major projects on the Karun river in the south-west, and a 1000 MW pumped storage plant at Siabisheh on the Caspian sea in the north

Reference summaries

Summary A

In Asia, China is planning to build the Three Gorges Dam on the Yangtze River and another on the Yellow River; India is building a dam on the Narmada River; Iran is planning four dams on the Karun River; Malaysia is planning a dam at Bakun in Sarawak on Borneo and another on the Pergau River; Laos is planning 58 dams on the Meking River; Nepal is planning a dam on the Arun River; and Turkey is planning the 22-dam Gap Project, the centerpiece of which is the Ataturk dam under construction

In Africa, Kenya is planning three dams on the Ewaso Ngiro River

In Europe, Portugal is planning a dam at Vila Nova de Foz Coa, and Slovakia is planning a dam on the Danube River

In the Western Hemisphere, Canada is planning a dam on the Quebec's Great Whale River; Mexico is planning a dam on Rio Usamacinta River; Chile is building the Pangue dam on the Bo-Bo River, one of six planned; and Paraguay together with Argentina are building the Yacyreta dam on the Paraha River and are planning another, the Corpus Cristi dam

Problems associated with the projects include the need to resettle people; environmental consequences, such as the destruction of rainforests; flooding issues, such as downriver flooding and flooding of important archeological sites; dependence on foreign capital, technologies and parts; and conflicts with neighboring states and countries because rivers flowing into them are being diverted, possibly affecting their power sources or causing silt accumulation

Table 6 (continued)

Summary B

New hydroelectric projects pose social, political, economic and environmental problems
China's Three Gorges project on the Yangtse would: displace 1.13 million people; do major social and environmental damage to Sichuan province while providing the most benefits to Hubei province; and be of enormous expense
In Malaysia The Pergau project pits its secular national government against Kilantan state's Islamist rulers
In Sarawak the Bakun project would displace 8000 tribespeople from rainforest which could become an "ecological time bomb"
The Narnada River project in northeastern India divides the three states involved and endangers the environment
The Arun River project in northern Nepal raises questions of costs versus benefits
Disputes over sharing river water bedevil the Bos Gablikova project on the Danube (Slovakia vs. Hungary), the GAP project in southeastern Turkey (Turkey vs. Syria and Iraq) and the Mekong River projects (Thailand, Vietnam, Cambodia and Laos)
Iran's Karun River project needs foreign financing as does Kenya's Evaso-Hgiro River project
Mexico's Rio Usumacinta project would flood ancient Mayan cities and threaten rainforest and its resident Indian tribe while the Pangue project on Chile's Bo-Bo River poses danger to a delicate ecosystem and the culture of the Pehuenche Indians
The Yacreta dam on the Parana River between Paraguay and Argentina caused years of political squabbles between the two countries
Paraguay needs private financing of its Corpus Cristi Dam while Canada's Great Whale River project faces opposition from environmentalists and Quebec's aboriginal community while there is reduced demand for electricity.

Summary C

A major obstacle for hydroelectric projects is opposition from environmental groups
Projects encountering this problem include China's Three Gorges project on the Yangtze River, Hydro-Quebec's Great Whale River project in Canada, Mexico's Rio Usumacinta dam, Chile's Pangue Project on the Bo-Bo River, Malaysia's Bakun project, India's Narmada River project, and Nepal's Arun project
Finance is another problem
Often, environmental opposition makes the World Bank reluctant to fund projects
High costs have delayed China's Three Gorges project, Kenya's Ewaso Ngiro project, Argentina's and Paraguay's Yacyreta dam, Nepal's Arun project, Turkey's south-east Anatolian project, and Iran's planned Karun River projects
Projects encounter opposition over the displacement of people or the destruction of cultural and historic resources
China's Three Gorges project would displace many people in Sichuan Province
Mexico postponed its Rio Usumacinta dam partly because of the possibility of flooding Mayan cities
Canada's Great Whale River project is opposed by the province's aboriginal community
Malaysia's Bakun project poses many technical and ecological challenges, as well as the possibility of damaging a unique tribal culture
Political issues plague many projects
The Yacyreta dam was delayed by Argentina's numerous political and economic crises
Malaysia's Pergau project is opposed by the state government of Kelantan
Hungary considers Slovakia's attempt to divert the River Danube to its own Cunovo project a territorial violation
Turkey's south-east Anatolian project poses political problems with Syria and Iraq
Projects which haven't encountered such difficulties include Portugal's Vila Nova de Foz Coa dam, China's Xiaolangdi project, and Paraguay's Corpus Cristi dam.

Summary D

There are many proposed hydroelectric projects being undertaken throughout the world
Each seems to be encountering problems of a financial, humanitarian, environmental, or political nature
Dam and hydroelectric projects are currently proposed for China's Yangtze River, in Kenya, in Mexico, in Quebec, in Southern Turkey, in Malaysia, in Slovakia on the Danube River, in Chile, in India on the Narmada River, in Iran, on the Parana River bordering Argentina and Paraguay, in Nepal, in Portugal, and in the Mekong Delta by Thailand, Vietnam, Cambodia, and Laos
Ecologists and environmentalists have objected to most of these projects because they will greatly alter the existing ecosystem
In Vietnam, the spawning grounds of 90% of the region's fish could be destroyed
Environmentalists also foresee destruction of forests as a problem, particularly in Chile
Environmentalists claim that some cases, such as the Mekong project, danger of serious flooding could increase
Projects in China, Malaysia, Quebec, and India would all result in lose of homeland of indigenous populations
The Mexican project would destroy archeological sites
Several dam projects have created serious controversies between nations
Hungary has objected to Slovakia's diverting of the Danube River and is threatening reprisals if the project goes through
Financial problems have plagued most of the dam sites
Kenya's project has led to questions about payments to consultants for study of the proposal
Turkey's economy has been strained by the massive costs of its hydroelectric project with many economists believing that it is the cause of Turkey's 70% rate of inflation

References

Bowman SR, Angeli G, Potts C, Manning CD (2015) A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on empirical methods in natural language processing, pp 632–642, https://doi.org/10.18653/v1/D15-1075

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al. (2020) Language models are few-shot learners. arXiv preprint arXiv:200514165

Canhasi E, Kononenko I (2014) Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization. Expert Syst Appl 41(2):535–543

Cao Z, Li W, Li S, Wei F, Li Y (2016) AttSum: Joint learning of focusing and summarization with neural attention. In: Proceedings of COLING 2016, the 26th International Conference on computational linguistics: Technical Papers, pp 547–556

Carbonell J, Goldstein J (1998) The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on research and development in information retrieval, pp 335–336

Celikyilmaz A, Hakkani RD (2010) A hybrid hierarchical model for multi-document summarization. In: Proceedings of the 48th Annual Meeting of the Association for computational linguistics, pp 815–824

Cer D, Yang Y, Kong Sy, Hua N, Limtiaco N, John RS, Constant N, Guajardo-Cespedes M, Yuan S, Tar C, et al. (2018) Universal sentence encoder. arXiv preprint arXiv:180311175

Conneau A, Kiela D, Schwenk H, Barrault L, Bordes A (2017) Supervised learning of universal sentence representations from natural language inference data. In: Proceedings of the 2017 Conference on empirical methods in natural language processing, pp 670–680, https://doi.org/10.18653/v1/D17-1070

Conroy JM, Schlesinger JD, Stewart JG (2005) Classy query-based multi-document summarization. In: Proceedings of the Document Understanding Conference

Daumé III H, Marcu D (2006) 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, pp 305–312

Devlin J, Chang M, Lee K, Toutanova K (2019) 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, NAACL-HLT 2019, pp 4171–4186

Dietterich TG (1998) Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput 10(7):1895–1923

Eheela J, Janet B (2020) An abstractive summary generation system for customer reviews and news article using deep learning. J Ambient Intell Hum Comput. https://doi.org/10.1007/s12652-020-02412-1

Ethayarajh K (2018) Unsupervised random walk sentence embeddings: a strong but simple baseline. In: Proceedings of The Third Workshop on representation learning for NLP, pp 91–100

Fabbri A, Li I, She T, Li S, Radev D (2019) 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, pp 1074–1084, https://doi.org/10.18653/v1/P19-1102

Feigenblat G, Roitman H, Boni O, Konopnicki D (2017) 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, pp 961–964

Haghighi A, Vanderwende L (2009) 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, pp 362–370

Howard J, Ruder S (2018) Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for computational linguistics (Volume 1: Long Papers), pp 328–339, https://doi.org/10.18653/v1/P18-1031

Iyyer M, Manjunatha V, Boyd-Graber J, Daumé III H (2015) Deep unordered composition rivals syntactic methods for text classification. In: Proceedings of the 53rd Annual Meeting of the association for computational linguistics and the 7th International Joint Conference on natural language processing, pp 1681–1691

Jain A, Bhatia D, Thakur MK (2017) Extractive text summarization using word vector embedding. In: 2017 International Conference on machine learning and data science (MLDS), pp 51–55

Joshi A, Fidalgo E, Alegre E, Fernández-Robles L (2019) Summcoder: an unsupervised framework for extractive text summarization based on deep auto-encoders. Expert Syst Appl 129:200–215

Kiros R, Zhu Y, Salakhutdinov RR, Zemel R, Urtasun R, Torralba A, Fidler S (2015) Skip-thought vectors. In: Cortes C, Lee DD, Sugiyama M, Garnett R (eds) Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, MIT Press, Montreal, Canada

Kobayashi H, Noguchi M, Yatsuka T (2015) Summarization based on embedding distributions. In: Proceedings of the 2015 Conference on empirical methods in natural language processing, EMNLP 2015, pp 1984–1989

Lamsiyah S, Mahdaouy AE, Espinasse B, Alaoui SOE (2020) An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2020.114152

Lebanoff L, Song K, Liu F (2018) Adapting the neural encoder-decoder framework from single to multi-document summarization. In: Proceedings of the 2018 Conference on empirical methods in natural language processing, pp 4131–4141, https://doi.org/10.18653/v1/D18-1446

Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L (2020) BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for computational linguistics, pp 7871–7880, https://doi.org/10.18653/v1/2020.acl-main.703

Lin CY (2004) Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out, Association for computational linguistics, Barcelona, Spain, pp 74–81

Liu Y, Lapata M (2019) Text summarization with pretrained encoders. 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), pp 3730–3740

Ma S, Deng ZH, Yang Y (2016) An unsupervised multi-document summarization framework based on neural document model. In: Proceedings of COLING 2016, the 26th International Conference on computational linguistics: Technical Papers, pp 1514–1523

Mao Y, Qu Y, Xie Y, Ren X, Han J (2020) Multi-document summarization with maximal marginal relevance-guided reinforcement learning. In: Proceedings of the 2020 Conference on empirical methods in natural language processing, EMNLP 2020, Online, November 16-20, 2020, pp 1737–1751, https://doi.org/10.18653/v1/2020.emnlp-main.136

Nenkova A, McKeown K (2011) Automatic summarization. Found Trends® Inf Retrieval 5:103–233. https://doi.org/10.1561/1500000015

Nenkova A, McKeown K (2012) A survey of text summarization techniques. In: Aggarwal, Charu C (eds) Mining text data, Springer US, Boston, MA, pp 43–76. https://doi.org/10.1007/978-1-4614-3223-4_3

Ouyang Y, Li W, Li S, Lu Q (2011) Applying regression models to query-focused multi-document summarization. Inf Process Manag 47(2):227–237

Pennington J, Socher R, Manning C (2014) GloVe: global vectors for word representation. In: Proceedings of the 2014 Conference on empirical methods in natural language processing (EMNLP), pp 1532–1543, https://doi.org/10.3115/v1/D14-1162

Radev DR, Jing H, Styś M, Tam D (2004) Centroid-based summarization of multiple documents. Inf Process Manag 40(6):919–938

Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf?fbclid=IwAR1N9fMDU7Txt2Sv0Vw6e3TVtLY75qSKfJbPP6NfdVrwzJsl49B80dlJvk

Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2019) Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:191010683

Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on empirical methods in natural language processing, pp 2383–2392, https://doi.org/10.18653/v1/D16-1264

Ramos J (2003) Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning, Piscataway, NJ, USA 242:133–142

Ren P, Chen Z, Ren Z, Wei F, Ma J, de Rijke M (2017) 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, pp 95–104

Ren P, Chen Z, Ren Z, Wei F, Nie L, Ma J, De Rijke M (2018) Sentence relations for extractive summarization with deep neural networks. ACM Trans Inf Syst (TOIS) 36:1–32

Robertson SE, Walker S, Jones S, Hancock-Beaulieu MM, Gatford M (1995) Okapi at trec-3. In: Overview of the Third Text REtrieval Conference (TREC-3), Gaithersburg, MD: NIST, pp 109–126. https://www.microsoft.com/en-us/research/publication/okapi-at-trec-3/

Roitman H, Feigenblat G, Cohen D, Boni O, Konopnicki D (2020) 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, pp 2577–2584, https://doi.org/10.1145/3dblp366423.3380009

Rossiello G, Basile P, Semeraro G (2017) Centroid-based text summarization through compositionality of word embeddings. In: Proceedings of the MultiLing 2017 Workshop on summarization and summary evaluation across source types and genres, pp 12–21

Ruder S, Peters ME, Swayamdipta S, Wolf T (2019) Transfer learning in natural language processing. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for computational linguistics: tutorials, pp 15–18

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252

Sakai T, et al. (2019) 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), IEEE, pp 153–157

Shen C, Li T, Ding CH (2011) 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

Valizadeh M, Brazdil P (2015) Exploring actor-object relationships for query-focused multi-document summarization. Soft Comput 19(11):3109–3121

Van Lierde H, Chow TW (2019a) Learning with fuzzy hypergraphs: a topical approach to query-oriented text summarization. Inf Sci 496:212–224

Van Lierde H, Chow TW (2019b) Query-oriented text summarization based on hypergraph transversals. Inf Process Manag 56(4):1317–1338

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, Curran Associates, Inc, Long Beach, CA, USA

Wan X, Xiao J (2009) Graph-based multi-modality learning for topic-focused multi-document summarization. In: Twenty-First International Joint Conference on artificial intelligence

Wan X, Zhang J (2014) Ctsum: extracting more certain summaries for news articles. In: Proceedings of the 37th International ACM SIGIR Conference on research & development in information retrieval, pp 787–796

Wieting J, Gimpel K (2018) ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In: Proceedings of the 56th Annual Meeting of the Association for computational linguistics (Volume 1: Long Papers), pp 451–462, https://doi.org/10.18653/v1/P18-1042

Wu Y, Li Y, Xu Y (2019) Dual pattern-enhanced representations model for query-focused multi-document summarisation. Knowl-Based Syst 163:736–748

Xiong S, Ji D (2016) Query-focused multi-document summarization using hypergraph-based ranking. Inf Process Manag 52(4):670–681

Xu Y, Lapata M (2020) Coarse-to-fine query focused multi-document summarization. In: Proceedings of the 2020 Conference on empirical methods in natural language processing (EMNLP), pp 3632–3645, https://doi.org/10.18653/v1/2020.emnlp-main.296

Yao Jg, Wan X, Xiao J (2015) Compressive document summarization via sparse optimization. In: Twenty-Fourth International Joint Conference on artificial intelligence

Yousefi-Azar M, Hamey L (2017) Text summarization using unsupervised deep learning. Expert Syst Appl 68:93–105

Zhong M, Liu P, Chen Y, Wang D, Qiu X, Huang X (2020) Extractive summarization as text matching. In: Jurafsky D, Chai J, Schluter N, Tetreault JR (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pp 6197–6208, doi 10.18653/v1/2020.acl-main.552

Zhong Sh, Liu Y, Li B, Long J (2015) Query-oriented unsupervised multi-document summarization via deep learning model. Expert Syst Appl 42(21):8146–8155

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com