A Comparative Study of Sentence Embeddings for Unsupervised Extractive Multi-Document Summarization

Salima Lamsiyah and Christoph Schommer

University of Luxembourg, Department of Computer Science, Luxembourg
{salima.lamsiyah,christoph.schommer}@uni.lu

Abstract. Obtaining large-scale and high-quality training data for multi-document summarization (MDS) tasks is time-consuming and resource-intensive, hence, supervised models can only be applied to limited domains and languages. In this paper, we introduce unsupervised extractive methods for both generic and query-focused MDS tasks, intending to produce a relevant summary from a collection of documents without using labeled training data or domain knowledge. More specifically, we leverage the potential of transfer learning from recent sentence embedding models to encode the input documents into rich semantic representations. Moreover, we use a coreference resolution system to resolve the broken pronominal coreference expressions in the generated summaries, aiming to improve their cohesion and textual quality. Furthermore, we provide a comparative analysis of several existing sentence embedding models in the context of unsupervised extractive multi-document summarization. Experiments on the standard DUC'2004-2007 datasets demonstrate that the proposed methods are competitive with previous unsupervised methods and are even comparable to recent supervised deep learning-based methods. The empirical results also show that the SimCSE embedding model, based on contrastive learning, achieves substantial improvements over strong sentence embedding models. Finally, the newly involved coreference resolution method is proven to bring a noticeable improvement to the unsupervised extractive MDS task.

Keywords: Unsupervised Multi-Document Summarization · Sentence Embeddings · Transfer Learning · Contrastive Learning · Coreference Resolution.

1 Introduction

Automatic Text Summarization (ATS) is the task of automatically condensing long documents into a shorter version that covers the main themes of those documents. There are two main approaches for ATS: *extractive approach* and *abstractive approach*. In the former, summaries are produced by identifying and extracting the most relevant sentences from the source documents, while in the latter, summaries are generated by reformulating and fusing ideas and often

by using a new lexicon. Abstractive methods may produce coherent and less redundant summaries based on natural language generation, while extractive methods enjoy better factuality and efficiency [4]. Motivated by the latter, we propose to improve the extractive approach by incorporating text understanding methods as well as coreference resolution techniques.

More precisely, we focus on multi-document summarization (MDS) that aims to produce a summary from a collection of thematically related documents. We consider both *generic* (G-MDS) and *query-focused* (QF-MDS) tasks. G-MDS systems produce summaries that represent all relevant facts of the source documents without considering the users' information needs. Besides, QF-MDS systems generate summaries that answer specific users' queries [19, 20]. Furthermore, we adopt an *unsupervised approach* that does not require labeled training data nor domain knowledge. Multi-document summarization task has received significantly less attention compared to single-document summarization, partly due to the scarcity of suitable data required for learning models [23]. Human annotation for summarization tasks, especially MDS, is a substantial time-requiring and costly manual effort. It is also unrealistic to expect that large-scale and high-quality labeled datasets will be created for different styles, domains, and languages. Additionally, introducing multi-document into the summarization task causes other difficulties. For instance, the extracted sentences may contradict each other because there is more diverse and conflicting information among documents. Moreover, information redundancy is omnipresent in MDS and has a significant impact on the information diversity of the generated summaries. The complexities of all these issues make the multi-document summarization a challenging task.

Furthermore, generating a relevant summary is a cognitive process that requires a deep understanding of the source documents as well as linguistic competence. Thus, creating internal representation to understand and analyze the semantic information of the source documents is a cornerstone step in text summarization methods. Bag-of-words and word embedding representations have shown promising results in text summarization [28, 36]. However, they do not consider the ordering of words in sentences as well as the semantic and syntactic relationships between them, and thus they may map semantically similar sentences into different vectors. Therefore, we need more accurate text representation methods that capture the semantic content of the source documents.

Recently, contextual pre-trained sentence embedding models, including inferSent [9], BERT encoder [10], simCSE [16], sentence-BERT [32], and others have demonstrated impressive performance in various NLP tasks [9, 10, 16, 32]. In this work, we apply several existing sentence embedding models to represent the documents' sentences as dense vectors in a low dimensional vector space and determine how well they capture relevant information to the unsupervised extractive multi-document summarization tasks (G-MDS and QF-MDS). Furthermore, we assess their performance, using ROUGE method [22], on the standard DUC'2003-2004 and DUC'2005-2007 datasets for the G-MDS and QF-MDS tasks, respectively. The experimental results show that the simCSE embedding

model [16], based on contrastive learning [8], brings substantial improvement over several other strong sentence embedding models.

Meanwhile, despite countless successes of the extractive methods, the generated summaries may contain incoherent sentences, as pronominal coreference expressions may appear unbound [1]. To alleviate this issue, we use a coreference resolution system (i.e. NeuralCoref[1]) to detect the broken pronominal coreference expressions in the selected sentences, and then rewrite those sentences by substituting with correct mentions. Advantageously, the proposed methods have achieved encouraging results, as the final summaries reached better ROUGE scores. We find also that our unsupervised extractive methods (G-MDS and QF-MDS) yield promising performance compared to the best-performing systems, including recent supervised deep learning-based methods.

The paper consists of the following sections: In Section 2, we briefly review the recent existing sentence embedding models. In Section 3, we present our generic and query-focused multi-document summarization methods. In Section 4, we analyze and compare the strengths and weaknesses of the described models and methods. Finally, in Section 5, we conclude the paper and outline some future directions in the field.

2 Related Work

The main objective of this work is to assess the performance of recent sentence embeddings in the context of unsupervised extractive multi-document summarization, considering both *generic* and *query-focused* tasks. To make the paper self-contained for reading, we briefly introduce the sentence embedding models exploited in this work. However, for readers who are interested in an overview of text summarization methods, they may refer to these recent surveys [12, 15].

Several sentence embedding methods exist that aim to encode sentences into dense vectors, which accurately capture the semantic and syntactic relationships between these sentences' constituents. Early work mostly concentrate on weighting and averaging words embedding vectors to construct the sentence embedding vector. In this context, the author in [13] has introduced the unsupervised smoothed inverse frequency (uSIF) model, which uses a pre-trained word vector model, tuned on the ParaNMT-50 dataset [38], 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.

In recent years, learning universal sentence embeddings using pre-trained models has gained much attention in NLP and tackled extensively in the literature [7, 9, 10, 16, 32]. For instance, Cer et al. [7] have introduced the universal sentence encoder DAN (USE-DAN) based on a deep average network [17]. It takes the average of word embeddings and bi-grams as input, which are then passed through a feed-forward neural network to produce the final sentence embedding vector. It is trained using unlabeled data selected from Wikipedia, web

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

news, web question-answer pages and discussion forums. Then, it is fine-tuned on the natural language inference task using the Stanford Natural Language Inference (SNLI) dataset [2], which has shown promising performance in various NLP tasks [7]. Besides, other embedding models use recurrent neural networks for learning universal sentences' representations. In this context, the supervised InferSent [9] trains a bi-directional long short-term memory network with max-pooling on the SNLI dataset. It has proven the suitability of natural language inference for transfer learning to other NLP tasks.

Furthermore, pre-trained sentence embedding models based on the Transformer architecture [37] have shown tremendous success in text encoding [7, 10, 16, 32]. In this context, Cer et al. [7] have introduced the universal sentence encoder (USE-Transformer) that uses the encoding sub-graph of the Transformer for sentence representation learning. Similarly to the USE-DAN model, the USE-Transformer is also trained on unlabeled data from Wikipedia, web news, web question-answer pages and discussion forums. Then, it is fine-tuned on the SNLI and the question-answering SQuAD datasets [2, 29]. In the same vein, other researchers have introduced the Bidirectional Encoder Representations from Transformers (BERT) model [10], which is based on a multi-layer bidirectional transformer encoder with attention mechanisms. BERT is trained on a large amount of unlabeled data selected from English Wikipedia and BookCorpus, using two unsupervised tasks: the masked language modeling and the next sentence prediction. Then, the pre-trained BERT can be fine-tuned on new NLP tasks using task-specific data. It has achieved impressive performance in a wide range of NLP tasks, including single text summarization [24].

Nevertheless, other researchers have introduced Sentence-BERT (SBERT) model [32], a modified version of the original pre-trained BERT model, which is mainly based on the siamese neural networks [3]. SBERT combines two BERT encoders into a siamese architecture to process two sentences in the same way, simultaneously. This two sub-networks derive semantically meaningful sentence embeddings[2], which can be then compared using the cosine similarity metric. It is trained on the combination of the SNLI [2] and Multi-Genre NLI [39] using the *classification*, *regression*, and *triplet* objective functions, depending on the available training data. Indeed, SBERT has shown state-of-the-art performance on the common STS benchmark [6] and transfer learning tasks.

More recently, the SimCSE embedding model [16], based on contrastive learning [8], has greatly advanced state-of-the-art sentence embedding methods. Contrastive Learning [8] is a machine learning paradigm that aims to learn effective representation by pushing semantically close neighbors towards each other in the embedding space, while pulling non-neighbors against each other. In conjunction with this, Gao et al. [16] have introduced two variants of SimCSE: 1) The *unsupervised SimCSE* that simply takes an input sentence and predicts itself in a contrastive learning framework, with only standard dropout used as noise; 2) The *supervised SimCSE* that incorporates annotated pairs from the NLI datasets [2, 39] into contrastive learning by using entailment pairs as posi-

[2] *Semantically meaningful* means similar sentences are close in the vector space.

tives and contradiction pairs as hard negatives. The authors have demonstrated that the contrastive learning objective can be extremely effective when coupled with pre-trained language models such as BERT [10].

Sentence embedding models are often evaluated at the time of their introduction with regard to the current state-of-the-art methods. Hence, it is rare when a single work presents a comparison of several embedding models for a specific task, we need to gather the results from numerous individual contributions. To the best of our knowledge, this is the first work that presents a comparative study of sentences embedding models for unsupervised generic and query-focused multi-document summarization tasks.

3 Multi-Document Summarization Methods

An extractive multi-document summarization method aims to identify and extract the most relevant sentences from a cluster of documents and adequately assemble them to form the final summary. Generally, the process of an extractive method involves the following main steps: text pre-processing, text representation, sentence scoring and selection, and sometimes a few extractive methods include other sentence-level operations such as sentence reordering or coreference resolution as a post-processing step [1].

3.1 Text Pre-processing

Given a cluster $D = \{d_1, d_2, ..., d_n\}$ consisting of n documents, we first split each document d_i into a set of sentences using the spaCy library[3], in particular, the pre-trained model "en_core_web_md". Then, we use the NLTK library[4] and regular expressions to perform tokenization, lowercasing, stemming, and to remove special characters (e.g. XML/HTML tags, URLs, email addresses, and redundant white-space). Hence, we obtain a cluster D of N sentences, denoted as $D = \{S_1, S_2, ..., S_N\}$. It is worth mentioning that for the query-focused summarization task, we also need to pre-process the pre-given user's query and represent it as a simple sentence Q.

3.2 Text Representation

Text representation plays a central role in extractive multi-document summarization methods to understand the content of the source documents. Thus, we leverage the potential of sentence embedding models (described in Section 2) to convert the documents' sentences into numeric fixed-length vectors that capture their semantic. They are two main approaches to use pre-trained sentence embedding models, namely 1) *feature-based approach* and 2) *fine-tuning approach*. In the former, the pre-trained model is used to extract fixed features for the input

[3] https://spacy.io/
[4] https://www.nltk.org/

documents' sentences, which can be used as input to the task at hand without any other modification. In the latter, the pre-trained model is fine-tuned on the downstream task where parameters of some layers are fixed and others are learned using the task-specific data.

Since we introduce in this paper *unsupervised* extractive methods, we opt for the first approach (*feature-based approach*). More precisely, given the cluster $D = \{S_1, S_2, ..., S_N\}$, we use a sentence embedding model (e.g. BERT, SBERT, and others) to map each sentence S_i in D into an embedding vector $\overrightarrow{S_i^D}$. Note that for the QF-MDS task, we also map the user's query into an embedding vector $\overrightarrow{Q}$ using the same sentence embedding models.

3.3 Sentence Scoring and Selection

Sentence scoring methods assign a score for each sentence in the cluster of documents to decide which sentences are most relevant to be selected as summary. For the generic G-MDS task, we measure the relevance of each sentence S_i in the cluster D without taking account of the user's specific need, while for the query-focused QF-MDS task, we score each sentence S_i in D based on its relevance to the input user's query Q. Our G-MDS and QF-MDS sentence scoring methods are successively described in the following.

Generic-MDS Task. Given a cluster $D = \{S_1, S_2, ..., S_N\}$ of N sentences, we assign a score for each sentence S_i in D by linearly combining three metrics, namely sentence content relevance, sentence novelty, and sentence position.

- **Sentence content relevance score**, formally defined in Eq. 1, is computed using the cosine similarity between the sentence embedding vector $\overrightarrow{S_i^D}$ and the centroid embedding vector of the cluster of documents $\overrightarrow{C_D}$ (defined in Eq. 2).

$$score^{contRelevance}(S_i, D) = \frac{\overrightarrow{S_i^D}.\overrightarrow{C_D}}{|| \overrightarrow{S_i^D} || \, . \, || \overrightarrow{C_D} ||} \tag{1}$$

$$\overrightarrow{C_D} = \frac{1}{N} \sum_{i=1}^{N} \overrightarrow{S_i^D} \tag{2}$$

  Where $\overrightarrow{C_D}$ is the centroid embedding vector of the cluster D, N is the number of sentences in D, and $\overrightarrow{S_i^D}$ is the embedding vector of the sentence S_i. The $score^{contRelevance}$ is bounded in [0,1] where sentences with higher scores are considered more relevant.

- **Sentence novelty score**, denoted as $score^{novelty}(S_i, D)$, is explicitly used to deal with redundancy and to produce summaries with good information diversity [20]. More precisely, for each sentence S_i in the cluster D, we compute its similarity with all the other sentences in D using the cosine similarity

between their corresponding embedding vectors. Then, if the maximum of the obtained similarities is below a given threshold τ, then the sentence S_i is considered novel. However, when the similarity between two sentences is greater than the given threshold, the sentence with the higher content relevance score gets the higher novelty score.

– **Sentence position** assigns a score for each sentence based on its position in the document, assuming that the first sentences of a document are more relevant to the summary [20]. Given D a cluster of n documents where each document d consists of M sentences $d = \{S_1, S_2, ..., S_M\}$, we compute the sentence position score of each sentence S_j in d using the following equation:

$$score^{position}(S_j) = \max(0.5, \exp(\frac{-p(S_j)}{\sqrt[3]{M}})) \quad (3)$$

Where M is the number of sentences in the document d, and $-p(S_j)$ is the jth position of S_j in d with $p(S_j) = 1$ for the first sentence and so on. The $score^{position}(S_j)$ is bounded between 0.5 and 1, where it is higher for sentences located at the beginning of the document. It gets stable at a value of 0.5 after a given number of sentences depending on the total number of sentences in the document.

Finally, we linearly combine these three metrics to get the final score of each sentence S_i in the cluster D, formally defined in the following equation:

$$score^{final}(S_i, D) = \alpha * score^{contRelevance} + \beta * score^{novelty} + \lambda * score^{position} \quad (4)$$

Where, $\alpha + \beta + \lambda = 1$ with $\alpha, \beta, \lambda \in [0, 1]$ with constant steps of 0.1. The top-ranked sentences are iteratively selected to form the summary w.r.t. the constraint on summary length L.

Query-Focused-MDS Task. Given a cluster $D = \{S_1, S_2, ..., S_N\}$ of N sentences and a user's query Q, we measure the relevance of each sentence S_i in D according the query Q using the cosine similarity between their embedding vectors $\overrightarrow{S_i^D}$ and $\overrightarrow{Q}$, respectively. Then, based on the obtained scores, we iteratively select the top-k ranked sentences such as $k \in \{50, 100\}$, formally denoted as top-$k = \{S_1, S_2, ..., S_k\}$. Next, we use a modified Maximal Marginal Relevance method [5, 19] that incorporates sentence embeddings to re-rank the top-k selected sentences intending to produce summaries that are relevant to the query Q and less redundant.

Finally, based on the obtained MMR sentences' scores, we apply a greedy search algorithm to select the relevant sentences to the input user's query, where a new sentence is added to the current summary if the constraint on the length limit L is not reached and the semantic similarity between this sentence and the already selected summary sentences is below a threshold τ.

3.4 Post-processing

As previously stated, extractive methods have proven to be effective for text summarization tasks, however, the generated summaries may lack cohesiveness since they sometimes contain broken pronouns. To address this issue, we use the NeuralCoref[5] model to detect each broken pronoun in the generated summary and then replace it with its corresponding entity. Nevertheless, as illustrated in the following example $\mathbf{S}_1$, the simple strategy of replacing every pronoun may cause redundant information and repetitive entity references in the generated summary.

– **Example $\mathbf{S}_1$:** On primary and secondary education, **Mrs Gillian Shephard**, the education secretary, announced tougher standards for teaching English in England and Wales, **she (Mrs Gillian Shephard)** launched an initiative to raise public consciousness about the need for good communication skills

To deal with this issue, we apply a rule-based heuristic that for each sentence in the generated summary, it keeps the pronoun if it appears after its referents; otherwise, the pronoun is unbound and must be replaced by its entity. The idea is very straightforward: substituting only the anaphoric expressions whose contexts are not present in the produced summaries.

4 Experiments

In this work, we are more interested in the degree to which the different sentence embedding models capture contextual and relevant information for solving unsupervised multi-document summarization tasks. Therefore, we present in this section all the experiments that are carried out to investigate the performance of the exploited models w.r.t the G-MDS and QF-MDS tasks.

4.1 Evaluation Datasets and Metrics

The experiments are carried out using the standard DUC'2003-2007 benchmarks[6], created essentially for evaluating multi-document summarization tasks. Table 1 summarizes some basic statistics of the used datasets.

For the evaluation measures, we use ROUGE (Recall-Oriented Understudy for Gisting Evaluation) method [22], in particular ROUGE-N (R-1, R-2, R-4) and ROUGE-SU4, adopting the same ROUGE settings[7] used for evaluating multi-document summarization methods.

[5] https://github.com/huggingface/neuralcoref
[6] https://duc.nist.gov/data.html
[7] ROUGE-1.5.5 with parameters "-n 4 -m -l 100 -c 95 -r 1000 -f A -p 0.5 -t 0" (G-MDS), "-a -c 95 -m -n 2 -2 4 -u -p 0.5 -l 250" (QF-MDS)

Table 1. A description of DUC'2003-2007 datasets [19, 20]. *Num docs* is the number of docs in each cluster. *Sum length* is the number of words in gold summaries.

Dataset	Clusters	Num docs	Sentences	Queries	Sum Length	Task
DUC'2003	30	10	7691	–	100	G-MDS
DUC'2004	50	10	13135	–	100	G-MDS
DUC'2005	50	32	45931	50	250	QF-MDS
DUC'2007	45	25	24282	45	250	QF-MDS

4.2 Experimental Setup

The introduced G-MDS and QF-MDS methods have been developed using PyTorch and a set of python tools, including the TrecTools[8] library and the available implementation of sentenc embedding models in Hugging Face[9], TensorFlow Hub[10], and GitHub[11]. Each model is designed to embed a sentence into a fixed dimensional length vector. For BERT, SBERT, and SimCSE-BERT, we used BERT_{base} model that produces embeddings vectors of 712 dimensions. The universal sentence encoders USE-DAN and USE-Transformer generate embeddings vectors of 512 dimensions, while the supervised InferSent-GloVe model produces embeddings vectors of 4090 dimensions.

For the G-MDS task, we need to optimize the hyper-parameters α, β, λ, and the threshold τ. Thus, we built a small held-out set by shuffling and randomly sampling 20 clusters from DUC'2002 dataset. Then, we performed a grid search on the held-out set under the condition $\alpha+\beta+\lambda = 1$, which gave us a total of 330 feasible combinations. Accordingly, the obtained values of the hyperparameters are 0.6, 0.2, 0.2, and 0.95 for α, β, λ, and τ, respectively.
For the QF-MDS task, we follow the same approach to optimize the three used hyper-parameters (i.e. the number of top ranked sentences k, the interpolation coefficient λ, and the threshold τ). We create a small held-out set by shuffling and randomly sample 20 clusters from DUC'2006 dataset. Then, we apply a grid search on the held-out set that gave us a total of 200 feasible combinations. Accordingly, the optimized values of λ, τ, and k are 0.9, 0.85, and 50, respectively.

Furthermore, for the statistical significance test, we used the *paired t-test* [11] to determine whether there is a significant difference in performance among all the evaluated models. Our choice is motivated by the fact that the authors in [31] have demonstrated that the *paired t-test* is more powerful than the equivalent unpaired test when applied to compare the outputs of two automatic text summarization systems. We attached a superscript to the performance number in the tables when the $p - value < 0.05$.

[8] https://pypi.org/project/trectools/
[9] https://huggingface.co/
[10] https://tfhub.dev/google
[11] https://github.com/facebookresearch/InferSent, https://github.com/kawine/usif

4.3 Results

Comparison of the different sentence embeddings w.r.t to the unsupervised G-MDS and QF-MDS tasks. The main objective of this paper is to examine the influence of transfer learning from sentence embedding models on the unsupervised multi-document summarization, considering both G-MDS and QF-MDS tasks. Tables 2 and 3 summarize the evaluation results of the different used text representation methods on G-MDS and QF-MDS tasks, including: a) Bag-of-words representation based on TF-IDF weighting scheme [30]; b) Word embeddings using the average of GloVe embeddings [27]; c) Unsupervised sentence embedding models using the average of BERT embeddings [10] and uSIF model [13]; d) Semi-supervised models using the universal sentence encoders (USE-Transformer, USE-DAN) [7]; e) Finally, supervised sentence embedding methods using InferSent [9], Sentence-BERT [32], and SimCSE [16] models.

Table 2. Comparison results of the used sentence embeddings w.r.t to the **G-MDS** task based on ROUGE recall scores. The superscripts *number* indicates significant improvement ($p-\text{value} < 0.05$) over the sentence embedding model that has the same superscript *number* attached.

MODELS	DUC'2003			DUC'2004		
	R-1	R-2	R-4	R-1	R-2	R-4
BOW and Word Embedding Models						
TF-IDF[1]	35.83^{3}	7.62^{3}	1.01^{3}	36.41^{3}	7.97^{3}	1.21^{3}
Avg. GloVe Embedding[2]	$36.72^{1,3}$	$8.45^{1,3}$	1.12^{3}	$37.10^{1,3}$	$8.80^{1,3}$	1.32^{3}
Unsupervised Models						
BERT Embedding[3]	28.03	4.48	0.45	28.92	4,43	0.59
uSIF[4]	38.29^{1-3}	9.27^{1-3}	$1.48^{1,3}$	39.72^{1-3}	9.79^{1-3}	1.65^{1-3}
Semi-Supervised Models						
USE-DAN[5]	38.35^{1-3}	9.06^{1-3}	$1.28^{1,3}$	$40.14^{1-3,7}$	9.85^{1-3}	1.58^{1-3}
USE-Transformer[6]	$38.56^{1-3,7}$	9.36^{1-3}	1.52^{1-3}	$40.32^{1-3,7}$	9.94^{1-3}	1.67^{1-3}
Supervised Models						
InferSent-GloVe[7]	37.59^{1-3}	9.01^{1-3}	$1.47^{1,3}$	38.71^{1-3}	9.17^{1-3}	$1.38^{1,3}$
SBERT[8]	39.24^{1-7}	9.72^{1-3}	1.68^{1-3}	$40.58^{1-3,7}$	10.04^{1-3}	1.84^{1-3}
SimCSE[9]	40.32^{1-8}	9.98^{1-7}	$1.92^{1-3,5}$	40.96^{1-8}	10.23^{1-7}	1.96^{1-7}

As shown in Table 2, the average of GloVe embeddings has significantly outperformed the TF-IDF model and BERT embeddings on the two used datasets and for most of the evaluation measures (R-1, R-2, and R-4). Noticing that using BERT model adopting the *feature-based approach* leads to rather poor performance, which is worse than computing the average of GloVe embeddings and TF-IDF model. The results show also that uSIF, USE-DAN, and USE-Transformer models have shown promising performance and outperformed the

supervised InferSent-GloVe model for most of evaluation measures. However, the difference between them is not statistically significant. Moreover, the SBERT model, based on siamese architecture, has outperformed all the other models except the SimCSE-BERT model, which has shown the best performance for all the evaluation metrics and on the two used datasets.

Table 3. Comparison results of the used sentence embeddings w.r.t to the **QF-MDS** task based on ROUGE recall scores. The superscripts *number* indicates significant improvement ($p-\text{value} < 0.05$) over the sentence embedding model that has the same superscript *number* attached.

MODELS	DUC'2005			DUC'2007		
	R-1	R-2	R-SU4	R-1	R-2	R-SU4
BOW and Word Embedding Models						
TF-IDF[1]	35.02	7.25	13.16	36.32	9.22	13.88
Avg. GloVe Embedding[2]	$37.66^{1,3}$	$7.67^{1,3}$	$14.05^{1,3}$	$40.22^{1,3}$	9.62^{3}	$15.23^{1,3}$
Unsupervised Models						
BERT Embedding[3]	35.15	6.64	12.62	36.63	7.74	13.24
uSIF[4]	$37.81^{1,3}$	$7.68^{1,3}$	$14.31^{1,3}$	$41.54^{1\text{-}3}$	10.08^{3}	$17.05^{1\text{-}3}$
Semi-Supervised Models						
USE-DAN[5]	$38.55^{1\text{-}4}$	$7.62^{1,3}$	$14.76^{1\text{-}3}$	$42.54^{1\text{-}4}$	$10.41^{1\text{-}3}$	$17.84^{1\text{-}4}$
USE-Transformer[6]	$39.65^{1\text{-}4,7}$	$8.21^{1\text{-}4,7}$	$15.5^{1\text{-}4}, 7$	$43.37^{1\text{-}4,7}$	$11.10^{1\text{-}4,7}$	$18.11^{1\text{-}4,7}$
Supervised Models						
InferSent-GloVe[7]	$38.03^{1\text{-}3}$	7.75	$14.47^{1,3}$	$42.06^{1\text{-}3}$	9.96^{3}	$17.37^{1\text{-}3}$
SBERT[8]	$40.07^{1\text{-}5,7}$	$8.57^{1\text{-}5,7}$	$15.72^{1\text{-}5,7}$	$43.75^{1\text{-}5,7}$	$11.27^{1\text{-}5,7}$	$17.96^{1\text{-}4}$
SimCSE[9]	$40.92^{1\text{-}8}$	$8.70^{1\text{-}7}$	$16.19^{1\text{-}7}$	$44.23^{1\text{-}8}$	$12.06^{1\text{-}7}$	$18.65^{1\text{-}4,7}$

As regards the QF-MDS task, the trend is similar. From Table 3, it seems clear that the SBERT and SimCSE-BERT models have achieved the best performance and led to significant improvement over most other models for all the evaluation measures (R-1, R-2, R-SU4). Furthermore, the universal sentence encoder USE-Transformer model has achieved better results than the USE-DAN and significantly outperformed the other models on both DUC'2005-2007 datasets and for most evaluation measures.

Therefore, the overall comparison of the exploited sentence embedding models has drawn the same conclusions on all the used datasets with regard to both G-MDS and QF-MDS tasks. More precisely, directly using BERT with no fine-tuning provides sentence embeddings, which are not suitable for the unsupervised multi-document summarization tasks; they yield slightly worse results than all the other models, including bag-of-words and word embeddings representations. Moreover, even though the supervised InferSent model is based on bidirectional LSTM networks and trained on the human-labeled SNLI dataset [2], it achieves comparable performance to the unsupervised uSIF model. This proves the ef-

fectiveness of uSIF sentence embeddings for the unsupervised multi-document summarization tasks.

Furthermore, the universal sentence encoder USE-Transformer has shown better performance that the USE-DAN. This can be due to their different architectures as well as the used training datasets. In fact, the USE-Transformer is further fine-tuned on the question-answering SQuAD dataset [29], and thus the knowledge gained from this related learning task helps boost the performance of the unsupervised MDS task. Additionally, we find that SBERT, based on siamese architecture and fine-tuning mechanisms, performs better than all the other previous models. Noticing that the SimCSE embedding model, based on the contrastive learning [8], has shown the best performance for both G-MDS and QF-MDS tasks.

Comparison with state-of-the-art methods. To prove the effectiveness of the introduced methods, we compare their performance with the best performing supervised and unsupervised state-of-the-art systems. ROUGE recall scores of the different systems used for comparison are summarized in Tables 4 and 5. It is worth mentioning that for the state-of-the-art methods, we report the results presented in their corresponding papers on DUC'2004 and DUC'2007, considered as the most used datasets for evaluating extractive G-MDS and QF-MDS systems. However, for our methods, we report the results obtained using the **SimCSE** embedding model, which achieves the best performance.

Table 4. ROUGE score of the G-MDS systems on DUC'2004 dataset.

System	**R-1**	**R-2**	**R-4**
DPP [18]	**39.79**	9.62	1.57
ConceptBased_ILP [26]	38.65	**10.02**	**1.67**
PG-MMR [21]	**36.42**	**9.36**	−−
Hi-MAP [14]	35.78	8.9	−−
GMDS-SimCSE (ours)	**40.96***	**10.23***	**1.96***

The first set of analyses is performed to compare our **GMDS-SimCSE** method with unsupervised systems, including the DPP [18] and the Concept-Based_ILP [26], which are considered as the best-performing extractive generic MDS systems on DUC'2004 dataset. As shown in Table 4, our method has achieved the best performance for all the evaluation measures (R-1, R-2, and R-4). More precisely, it achieves an increment of 1.17% w.r.t the DPP system for the R-1 measure and an improvement of 0.21% and 0.29% w.r.t the Concept-Based_ILP system for R-2 and R-4, respectively.

To further investigate the effectiveness of the introduced method, we compare its performance with recent supervised systems, including the PG-MMR [21] and

the Hi-MAP [14] that are mainly based on the pointer-generator networks with the maximal marginal relevance method [5]. As depicted in Table 4, our method GMDS-SimCSE has shown better performance than both the PG-MMR and HI-MAP methods. In particular, it has achieved an increment of 4,54% and 0,87% with respect to PG-MMR for R-1 and R-2 respectively. This can be explained by the fact that both PG-MMR and HI-MAP methods have been trained on the CNN\DailyMail datasets, mainly created for single-document summarization, where documents are very short compared to a cluster of documents.

The second set of analyses is conducted to evaluate the performance of our query-focused multi-document summarization method (**QFMDS-SimCSE**) using DUC'2007 dataset. We compare it against the best-performing state-of-the-art methods, including 1) the unsupervised Dual-CES [35] and USE-Transformer-Sum [19] systems, and 2) the supervised CRSum-SF [33] and SRSum [34] systems, which are based on convolutional neural networks with attention mechanisms. Therefore, as shown in Table 5, in terms of R-1 and R-2 our method has achieved comparable performance to all the other systems except Dual-CES, which has yielded very high performance w.r.t R-1 score. This can be because the Dual-CES system better handles the tradeoff saliency and focus in the summarization process. However, in terms of R-SU4 evaluation measure, our method has achieved the best performances.

Table 5. ROUGE recall score of the QF-MDS systems on DUC'2007 dataset.

System	**R-1**	**R-2**	**R-SU4**
Dual-CES [35]	**46.02***	12.53	17.91
USE-Transformer-Sum [19]	43.54	11.42	18.54
CRSum-SF [33]	44.6	12.48	−−
SRSum [34]	45.01	**12.8***	−−
QFMDS-SimCSE (ours)	44.23	12.06	**18.65***

The overall obtained results show that our unsupervised extractive multi-document summarization methods, based on sentence embeddings and coreference resolution, achieve promising results for both generic and query-focused tasks. Moreover, they yield far better performance than the state-of-the-art methods that are based on bag-of-words and word embeddings representations. A concrete example of gold summary and our **QFMDS-SimCSE** system's output is presented in Table 6 in Appendix A. It can be seen from this example that replacing the broken pronoun "**They**" in sentence "S_3" by its entity "**Richard Roberts and Phillip Sharp**" has improved the cohesion of the generated summary. Furthermore, this example shows also that the produced summary is relevant to the input query.

5 Conclusion and Future Directions

Different from other natural language processing applications, text summarization is a challenging task that is highly subjective and dependent on the content. Determining the relevance of information included in documents requires a deep understanding of the source documents. Therefore, the main objective of this paper was to investigate the performance of deep understanding methods, namely recent sentence embedding models, on the unsupervised extractive multi-document summarization, considering both generic and query-focused tasks. The results have shown that models based on the Transformer architecture and fine-tuned on the NLI datasets lead to strong results compared to other models. This proves the effectiveness of transfer learning from pre-trained sentence embedding models, which allows benefiting from knowledge learned from other related natural language understanding tasks to improve the performance of the target task. Additionally, the results have also shown that the SimCSE embedding model, based on contrastive learning, has demonstrated substantial improvement in extractive unsupervised MDS task.

Furthermore, as previously mentioned, supervised multi-document summarization methods require high-quality labeled training data, which is of immense importance for the success of these methods. However, acquiring such data for MDS is a cumbersome task, especially for specific domains, where experts are required to annotate the data. Thus, the recent focus of deep learning research is to reduce the requirement for supervision in model training. In fact, fine-tuning deep pre-trained language models has set state-of-the-art performance on a wide range of NLP applications, however, their generalization performance drops under domain shift. To mitigate this issue, several s*elf-supervised learning* methods have been introduced in the literature, which are based for instance on *unsupervised domain adaptation* or *contrastive learning* approaches.

Unsupervised domain adaptation methods aim to generalize well on the target domain by learning from both labeled samples from the source domain and unlabeled samples from the target domain, while contrastive learning methods learn to contrast between pairs of similar and dissimilar data points [8]. These methods have shown impressive performance in several NLP tasks [25]; they facilitate data-efficient learning, especially when training data is not abundantly available. Motivated by these findings, we plan to investigate the potential of self-supervised learning methods, based on unsupervised domain adaptation and contrastive learning, to improve the *abstractive* multi-document summarization tasks.

Finally, we believe that our unsupervised G-MDS and QF-MDS methods that do not require labeled training data nor domain knowledge can be used as strong baselines for evaluating *extractive* multi-document summarization systems.

References

1. Antunes, J., Lins, R.D., Lima, R., Oliveira, H., Riss, M., Simske, S.J.: Automatic cohesive summarization with pronominal anaphora resolution. Computer Speech

& Language **52**, 141–164 (2018)
2. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: 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
3. Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. International Journal of Pattern Recognition and Artificial Intelligence **7**(04), 669–688 (1993)
4. Cao, Z., Wei, F., Li, W., Li, S.: Faithful to the original: Fact aware neural abstractive summarization. In: thirty-second AAAI conference on artificial intelligence (2018)
5. Carbonell, J., Goldstein, J.: 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 (1998)
6. Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 1–14 (2017)
7. Cer, D., Yang, Y., Kong, S.y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., Kurzweil, R.: Universal sentence encoder for English pp. 169–174 (2018)
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607 (2020)
9. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP. pp. 670–680 (2017)
10. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding pp. 4171–4186 (2019)
11. Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation **10**(7), 1895–1923 (1998)
12. El-Kassas, W.S., Salama, C.R., Rafea, A.A., Mohamed, H.K.: Automatic text summarization: A comprehensive survey. Expert Systems with Applications p. 113679 (2020)
13. Ethayarajh, K.: Unsupervised random walk sentence embeddings: A strong but simple baseline. In: Proceedings of The Third Workshop on Representation Learning for NLP. pp. 91–100 (2018)
14. Fabbri, A., Li, I., She, T., Li, S., Radev, D.: 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 (2019)
15. Gambhir, M., Gupta, V.: Recent automatic text summarization techniques: a survey. Artificial Intelligence Review **47**(1), 1–66 (2017)
16. Gao, T., Yao, X., Chen, D.: SimCSE: Simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 6894–6910 (2021)
17. Iyyer, M., Manjunatha, V., Boyd-Graber, J., Daumé III, H.: 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 (2015)
18. Kulesza, A., Taskar, B., et al.: Determinantal point processes for machine learning. Foundations and Trends® in Machine Learning **5**(2–3), 123–286 (2012)
19. Lamsiyah, S., El Mahdaouy, A., El Alaoui, S.O., Espinasse, B.: Unsupervised query-focused multi-document summarization based on transfer learning from sentence embedding models, bm25 model, and maximal marginal relevance criterion. Journal of Ambient Intelligence and Humanized Computing pp. 1–18 (2021)
20. Lamsiyah, S., El Mahdaouy, A., Espinasse, B., Ouatik, S.E.A.: An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings. Expert Systems with Applications **167**, 114152 (2021)
21. Lebanoff, L., Song, K., Liu, F.: 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 (2018)
22. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81 (2004)
23. Liu, Y., Lapata, M.: Hierarchical transformers for multi-document summarization. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5070–5081 (2019)
24. Liu, Y., Lapata, M.: 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 (2019)
25. Long, Q., Luo, T., Wang, W., Pan, S.: Domain confused contrastive learning for unsupervised domain adaptation. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2982–2995 (2022)
26. Oliveira, H., Lins, R.D., Lima, R., Freitas, F., Simske, S.J.: A concept-based ilp approach for multi-document summarization exploring centrality and position. In: 2018 7th Brazilian Conference on Intelligent Systems (BRACIS). pp. 37–42 (2018)
27. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
28. Radev, D.R., Jing, H., Styś, M., Tam, D.: Centroid-based summarization of multiple documents. Information Processing & Management **40**(6), 919–938 (2004)
29. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (2016)
30. Ramos, J., et al.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. vol. 242, pp. 29–48 (2003)
31. Rankel, P.A., Conroy, J., Slud, E., O'leary, D.P.: Ranking human and machine summarization systems. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 467–473 (2011)
32. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3982–3992. Hong Kong, China (Nov 2019)

33. Ren, P., Chen, Z., Ren, Z., Wei, F., Ma, J., de Rijke, M.: 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 (2017)
34. Ren, P., Chen, Z., Ren, Z., Wei, F., Nie, L., Ma, J., De Rijke, M.: Sentence relations for extractive summarization with deep neural networks. ACM Transactions on Information Systems (TOIS) **36**, 1–32 (2018)
35. Roitman, H., Feigenblat, G., Cohen, D., Boni, O., Konopnicki, D.: 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 (2020)
36. Rossiello, G., Basile, P., Semeraro, G.: 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 (2017)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
38. Wieting, J., Gimpel, K.: 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 (2018)
39. Williams, A., Nangia, N., Bowman, S.: A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1112–1122 (2018)

A Example of our QFMDS-SimCSE Output's Summary

Table 6. Example of the generated summary for **Cluster D374a** from DUC'2005 dataset using our **QFMDS-SimCSE** method.

Query
• $\mathbf{S}_1$ Who are the Nobel Prize winners in the sciences and in economics and what are their prize-winning achievements? What are common factors in their backgrounds?
Generated summary
• $\mathbf{S}_1$ THE WINNERS of the Nobel prizes, announced in Stockholm, made discoveries which helped to uncover some of the most fundamental processes in science.
• $\mathbf{S}_2$ The Nobel Prize in chemistry is shared by Thomas Cech, 41, of the University of Colorado, and Sidney Altman, 50, of Yale University.
• $\mathbf{S}_3$ **Richard Roberts and Phillip Sharp(They)** have just jointly been awarded the Nobel prize in medicine.
• $\mathbf{S}_4$ Jerome I. Friedman and Henry W. Kendall of the Massachusetts Institute of Technology and Canadian Richard E. Taylor of Stanford University will share the $700,000 Nobel Prize in physics.
• $\mathbf{S}_5$ Murray Gell-Mann won the Nobel prize for physics in 1969, and later helped establish the Santa Fe Institute, an interdisciplinary foundation devoted to the study of 'complex systems' as various as quantum mechanics, the human body, and international economics.
• $\mathbf{S}_6$ MR GARY BECKER, named as this year's winner of the Nobel prize for economics, is proof that economists have more to offer than dubious forecasts, indecipherable equations and contradictory conclusions about the behavior of money and markets.
• $\mathbf{S}_7$ THE Nobel prize for economics was awarded to Robert Fogel of the University of Chicago and Douglass North of Washington University in St Louis for pioneering work on the causes of economic and institutional change.
• $\mathbf{S}_8$ The Nobel Prize for Economics was awarded to three 'game theorists': John Harsanyi, John Nash and Rheinhard Selten.
• Norman Ramsey of Harvard University will receive half the physics prize for his discovery of the atomic clock.
• $\mathbf{S}_9$ The announcements completed a near-sweep of the science Nobel's by U.S. researchers this year, continuing U.S. dominance of the prizes.
Gold summary
• Nobel prizes are award each year for achievements in the physical sciences – physics, chemistry, medicine, economics, literature and for peace.
• Winners in physics include Norman Ramsey, Wolfgang Paul and Hans Delmelt in 1989 for work leading to the cesium atomic clock; Jerome Friedman, Henry Kendall and Richard Taylor in 1990 for first detecting quarks; Georges Charpak in 1992 for particle detectors; and Betran Brockhouse and Clifford Shull in 1994 for work on neutron scattering.
• Chemistry winners include Sidney Altman and Thomas Cech in 1989 for work on RNA; Elias Cory in 1990 for work on organic synthesis; and Rudolph Marcus in 1992 for electron transfer theory.
• Winners in medicine include J. Michael Bishop and Harold Varmus in 1989 for contributions to cancer research; Joseph Murray and E. Donnall Thomas in 1990 for work on organ transplants; Richard Roberts and Phillip Sharp in 1993 for cancer research; and Alfred Gilman and Martin Rodbell in 1994 for work on proteins.
• Among winners in economics are Robert Fogel and Douglas North in 1993 for work on causes of economic change; Gary Becker in 1992 for work on the economics of discrimination and human capital theory; and John Harsanyi, John Nash and Rheinhard Selten in 1994 for work on game theory.
• Winners share some common background factors, One is that generally took five and 20 years between a discovery and its recognition.
• Another is that most winners in certain fields were Americans – of 142 medicine prizes awarded, 69 were to Americans.