Journal Pre-proof

An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings

Salima Lamsiyah, Abdelkader El Mahdaouy, Bernard Espinasse, Saïd El Alaoui Ouatik

PII: S0957-4174(20)30895-2
DOI: https://doi.org/10.1016/j.eswa.2020.114152
Reference: ESWA 114152

To appear in: *Expert Systems With Applications*

Received date : 25 October 2019
Revised date : 3 September 2020
Accepted date : 22 October 2020

Please cite this article as: S. Lamsiyah, A. El Mahdaouy, B. Espinasse et al., An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings. *Expert Systems With Applications* (2020), doi: https://doi.org/10.1016/j.eswa.2020.114152.

An Unsupervised Method for Extractive Multi-Document Summarization based on Centroid Approach and Sentence Embeddings

Salima Lamsiyah[a,b,*], Abdelkader El Mahdaouy[b], Bernard Espinasse[c] and Saïd El Alaoui Ouatik[a,b]

[a]Laboratory of Engineering Sciences, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco.
[b]Laboratory of Informatics, Signals, Automatic, and Cognitivism, FSDM, Sidi Mohamed Ben Abdellah University, Fez, Morocco.
[c]Aix-Marseille Université, Université de Toulon, LIS UMR CNRS 7020, Marseille, France.

ARTICLE INFO

Keywords:
Extractive Text Summarization
Word Embeddings
Sentence Embeddings
Centroid Approach
Transfer Learning

ABSTRACT

Extractive multi-document summarization (MDS) is the process of automatically summarizing a collection of documents by ranking sentences according to their importance and informativeness. Text representation is a fundamental process that affects the effectiveness of many text summarization methods. Word embedding representations have shown to be effective for several Natural Language Processing (NLP) tasks including Automatic Text Summarization (ATS). However, most of these representations do not consider the order and the semantic relationships between words in a sentence. This does not fully allow grasping the sentence semantics and the syntactic relationships between sentences constituents. In this paper, to overcome this problem, we propose an unsupervised method for generic extractive multi-document summarization based on the sentence embedding representations and the centroid approach. The proposed method selects relevant sentences according to the final score obtained by combining three scores: sentence content relevance, sentence novelty, and sentence position scores. The sentence content relevance score is computed as the cosine similarity between the centroid embedding vector of the cluster of documents and the sentence embedding vectors. The sentence novelty metric is explicitly adopted to deal with redundancy. The sentence position metric assumes that the first sentences of a document are more relevant to the summary, and it assigns high scores to these sentences. Moreover, this paper provides a comparative analysis of nine sentence embedding models used to represent sentences as dense vectors in a low dimensional vector space in the context of extractive multi-document summarization. Experiments are performed on the standard DUC'2002-2004 benchmark datasets and the Multi-News dataset. The overall obtained results have shown that our method outperforms several state-of-the-art methods and achieves promising results compared to the best performing methods including supervised deep learning based methods.

1. Introduction

Nowadays, textual information in digital format is abundant on the web, and it represents about 80% of the information circulating there. Hence, we are facing an inevitable and challenging problem of information overload, which has highlighted the need to develop relevant and specific tools to alleviate this problem by allowing users to save time and resources, as well as to find the most suitable information for their needs. Automatic Text Summarization (ATS), by condensing texts while preserving their important aspects can help to process this ever-growing text collection efficiently. The main idea of automatic text summarization is to find a subset of data that contains the information of the entire set. Therefore, an ATS should deal with two fundamental issues (Saggion and Poibeau, 2013): (i) *how to select useful and relevant information*; and (ii) *how to express this information in a coherent and a concise form*. Different types of automatic text summarization have been proposed. For instance, regarding the number of input documents, single and multi-document summarization have been introduced. Similarly, based on the purpose of the summarization, it can be either generic or query-focused. Generic summaries represent all relevant facts of a source document

*Corresponding author
salimalamsiyah@gmail.com (S. Lamsiyah); abdelkader.elmahdaouy@usmba.ac.ma (A. El Mahdaouy); bernard.espinasse@lis-lab.fr (B. Espinasse); said.ouatikelalaoui@usmba.ac.ma (S. El Alaoui Ouatik)
ORCID(s):

without considering the users' information needs, whereas in query-focused summaries, the content of the summary is derived from the user's information need or simply the user's query. In this paper, we focus on generic extractive multi-document summarization.

Several methods have been proposed for automatic text summarization. Generally, they are classified into two categories: *extractive approach* and *abstractive approach*. The former is designed to identify and select the most relevant sentences exactly as they appear in the original documents. While the abstractive approach generates summaries by concisely paraphrasing the document's content, they are considered complex, and sometimes they are not completely automatic. They require resources previously built that demand a high computational effort. Thus, as part of our research, we will use the extractive approach for multi-document text summarization.

Extractive based methods consist mainly of three steps: (i) document analysis and representation; (ii) sentence scoring; and (iii) sentence selection. The first step preprocesses and analyzes the documents to build a representation of their content. Based on the latter representation, a score is assigned for each sentence to measure its relevance, and finally, the top-ranked sentences are selected to form the summary. A suitable extractive method must select the relevant sentences that satisfy and optimize coverage and diversity properties and also that minimizes the redundancy between the selected sentences.

Many extractive text summarization methods proposed in the literature are based on Bag-of-Words (BOW) representations of text documents (Radev et al., 2004). Despite their popularity, BOW features lose the ordering of words and ignore the semantics of the words. Even though bag-of-N-grams representations consider the words order in a short context, they suffer from data sparsity and the curse of dimensionality (Le and Mikolov, 2014). Based on the idea that words in similar contexts have a similar meaning, Kågebäck et al. (2014); Kobayashi et al. (2015a); Jain et al. (2017); Rossiello et al. (2017); Ghalandari (2017) have proposed to use word embedding methods which represent words as dense vectors in low-dimensional vector space using various pre-trained models inspired from neural networks language modeling. These representations have shown a better performance as a representational basis for ATS tasks. However, representing relationships among multiple words and phrases in a single dense vector is an emerging problem. For example, taking into account the following two sentences *"You are going there to study not to teach"* and *"You are going there to teach not to study"*, these two sentences will have an identical representation using word embeddings and BOW representations, while their meanings are entirely different.

Therefore, learning viable representations of input data has been considered as a hot topic in current natural language processing research. Indeed, the biggest challenge resides in learning the embedding vector of a whole sentence. Even though pre-trained word vectors like word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) have been able to capture the meaning of single words, they do not fully consider the contextual information and the structure of sentences. Recently, several sentence embedding models have been developed to learn distributed semantic representations of sentences. Most of these models are based on deep neural network architectures including recursive networks (Socher et al., 2013), recurrent networks (Hochreiter and Schmidhuber, 1997), convolutional networks (Kalchbrenner et al., 2014; Kim, 2014) and recursive-convolutional methods (Cho et al., 2014; Zhao et al., 2015) among others. In the past few years, pre-trained sentence embedding models have shown state-of-the-art results on a wide range of natural language processing tasks, including sentence similarity, sentiment analysis, question-answering, text classification, and many other tasks (Conneau et al., 2017; Kiros et al., 2015; Cer et al., 2018; Devlin et al., 2019). Sentence embedding methods aim to encode sentences into dense vectors, represented in a low dimensional vector space, that accurately capture the semantic and syntactic relationships between these sentences constituents. Thus, sentences with similar meanings are mapped to similar vectors, and simultaneously sentences with different meanings are mapped to different vectors. Moreover, many NLP researchers have demonstrated the importance of pre-trained sentence embedding models for transfer learning. Where these models are first pre-trained on a large amount of textual data, and then transfer the learned knowledge to another downstream NLP task. Hence, pre-trained sentence embedding methods allow us to save time and computational power as well as benefiting from knowledge learned across multiple related tasks. Following this success, we propose to improve the extractive multi-document summarization task by exploiting the capabilities of these representations.

In this paper, we propose an unsupervised extractive method for multi-document summarization based on the centroid approach and sentence embedding representations. Indeed, the centroid-based approach has already applied in multi-document text summarization (Radev et al., 2004) where the aim is to condense the meaningful information of a cluster of documents as a set of statistically important words, that could be used to identify the salient sentences. The centroid-based approach is often considered as a strong extractive multi-document summarization baseline in literature. However, other experiments have demonstrated that applying some modifications to this baseline can improve the

performance (Rossiello et al., 2017; Ghalandari, 2017). From this point of view, we proposed to improve this method by introducing two major modifications: First, instead of representing documents using BOW and word embedding representations, we explored the potential of the recent sentence embedding models. The latter choice is motivated by the fact that sentence embedding models are considered as a breakthrough in NLP and showed state-of-the-art performances for many NLP tasks, as well as the central role they play in transfer learning for NLP applications. Second, in order to produce summaries with more information diversity, we improved the sentence scoring function by combining linearly three metrics, namely sentence content relevance, sentence novelty, and sentence position.

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

1. We propose an unsupervised method for extractive multi-document summarization based on the centroid approach and the sentence embedding representations.
2. We improve sentence scoring by combining three metrics, including sentence content relevance, sentence novelty, and sentence position. We show that combining these three metrics leads to significant improvements.
3. We evaluate our method using nine sentence embedding models for extractive summarization task. We assess the performance of these models and then provide an empirical analysis of these representations on text summarization task using the standard DUC'2002-2004 datasets.
4. We compare our method with several state-of-the-art methods, including recent supervised deep learning based methods on DUC'2004 and Multi-News datasets. Experimental results show that our method has achieved comparable performances with several state-of-art methods.

We empirically evaluated the proposed method using ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics (Lin, 2004), more specifically, ROUGE-N (ROUGE-1, ROUGE-2, ROUGE-4), ROUGE-SU4, and ROUGE-L. The aim is to measure the content similarity between the generated summaries and the corresponding reference summaries (gold summaries). The obtained results show that the proposed method outperforms traditional baselines and yields comparable performances to the recent state-of-the-art methods for multi-document summarization including supervised deep learning based methods. Moreover, the proposed method is unsupervised, fast, and easy to implement, which can be used as a baseline for evaluating multi-document summarization systems.

The rest of this paper is organized as follows. We discuss the related work in Section 2. We present in detail the proposed method in Section 3. Section 4 describes the conducted experiments and presents the obtained results. While section 5 provides a discussion of the obtained results. Finally, Section 6 concludes this paper and draws lines for future work. The notations we use throughout the paper are summarized in Table 1.

2. Related Work

In this work, we propose an unsupervised method for generic extractive multi-document summarization based on the centroid approach and sentence embedding representations. For this reason, we will briefly present the recent sentence embedding models as well as some previous unsupervised extractive methods to make the paper self-contained for reading. For the reader who is interested in a detailed overview of automatic text summarization approaches and methods, he may refer to the recent surveys on the field (Yao et al., 2017; Aries et al., 2019).

2.1. Sentence Embedding Models

Traditional sentence embedding methods are based on weighting and averaging words vectors of their constituents to construct sentences' vectors. Recently, pre-trained sentence embedding models have emerged as important methods for learning contextual representations. Most of these models are pre-trained using language modeling tasks on large text corpora. The existing sentence embedding methods can be classified according to the learning paradigm into two categories: i) *parameterized methods* and ii) *non-parameterized methods*.

Parameterized sentence embedding methods require training to optimize their parameters. For instance, the probabilistic language model (Bengio et al., 2003), termed as NNLM, uses neural networks to model the conditional probability of a word, given a fixed number of the preceding words. Hence, the model learns simultaneously a distributed representation for each word as well as the probability function of word sequences, expressed in terms of representations. Skip-Thought (ST) (Kiros et al., 2015) is an unsupervised model for learning generic and distributed sentence representations called skip-Thought vectors. It is based on encoder-decoder models, where the encoder (usually based on Recurrent Neural Networks (RNNs)) maps words to a sentence vector, and the decoder predicts the surroundings

Table 1
Notations used in the paper

Notations	Description
d	Document of M sentences
M^d	Number of sentences of the document d
D	Cluster of m documents d
N	Number of sentences in D
S_i	The $i-th$ sentence in the cluster D
$\overrightarrow{S_i^D}$	Embedding vector of S_i in the cluster D
l	Index of the sentence most similar to S_i
$\overrightarrow{C_D}$	Centroid embedding vector of the cluster D
$p(S_i^d)$	Position of the $i-th$ sentence in the document d
τ	Similarity threshold between sentences
α	Weight of sentence content relevance
β	Weight of sentence novelty
λ	Weight of sentence position
$score^{contentRelevance}(S_i, D)$	Content relevance score of the $i-th$ sentence in the cluster D
$score^{novelty}(S_i, D)$	Novelty score of the of the $i-th$ sentence in the cluster D
$score^{position}(S_i^d)$	Position score of the $i-th$ sentence in the document d
$score^{final}(S_i, D)$	Final score of the $i-th$ sentence in the cluster D

sentences. Compared with a simple average of words embeddings representation, Skip-Thought model takes into account the order of words during the encoding/decoding process. In another work, Conneau et al. (2017) have proposed the InferSent Embedding Model as a supervised model for learning universal sentences' representations. The InferSent model is trained using the supervised data of the Stanford Natural Language Inference (SNLI) dataset (Bowman et al., 2015), considered as one of the largest and high-quality labeled corpus, designed for textual entailment tasks. The authors have exploited seven different architectures for encoding sentences into fixed-size representations, more precisely, the standard recurrent models such as Long Short Term Memory (LSTMs) and Gated Recurrent Units (GRUs). Therefore, they have shown that the bi-directional LSTM network with max-pooling trained with SNLI dataset (Bowman et al., 2015) makes the best sentence encoding method. Peters et al. (2018) have introduced the Embedding from Language Models (ELMo), a deep contextualized word representation that first models syntactic, semantic, and other complex characteristics of word use. Then, it captures how these uses vary across linguistic contexts. ELMo model uses a bi-directional LSTM, trained using language model (LM) on a massive text dataset. Hence, it provides a very rich representation of the token that can overcome the limitations of previous word embeddings where each word is usually modeled as an average of their multiple contexts. Moreover, Cer et al. (2018) have recently proposed two universal sentence encoders, namely the USE-DAN and USE-Transformer models. The USE-DAN uses a Deep Averaging Network (Iyyer et al., 2015), where embeddings of words and bi-grams are averaged together and then used as input to a feed-forward deep neural network to compute the sentence embedding. While the USE-Transformer model builds sentence embedding using the encoding sub-graph of the transformer architecture proposed by (Vaswani et al., 2017), which is solely based on attention mechanisms. This sub-graph computes context-aware representations of words in a sentence by taking into consideration both the ordering and the identity of all other words. Both the USE-DAN and USE-Transformer models take as input a lowercased Penn Treebank[1] (PTB) tokenized string and generate as output a 512-dimensional sentence embedding. Furthermore, Devlin et al. (2019) have proposed the Bidirectional Encoder Representations from Transformers (BERT), a recent text representation model trained on a large corpus of

[1] "https://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/process/PTBTokenizer.html"

3300 M words, with two unsupervised tasks namely the masked language modeling and the next sentence prediction. The BERT model is considered as a multi-layer bidirectional Transformer encoder that is based on the original Transformer's implementation described in (Vaswani et al., 2017). Where two versions of BERT have been trained, namely BERT_{BASE} and BERT_{LARGE}.

Non-parameterized sentence embedding methods do not require any external data and any training only pre-trained word embedding vectors. For instance, Ethayarajh (2018) proposes the Unsupervised Smooth Inverse Frequency (uSIF) sentence embedding model as a refinement to the Smooth Inverse Frequency (SIF) model (Arora et al., 2016). The author has shown that the word vector length has a confounding effect on the log-linear random walk model of generating sentences in SIF. Hence, he has proposed a random walk model that handles this confounds, in which the probability of word generation is inversely related to the angular distance between the word and the sentence embeddings. Thus, uSIF differs from SIF in that uSIF requires no hyper-parameter tuning, which means that it can be used when there is no labeled data, which makes it completely unsupervised. The Geometric Sentence Embedding (GEM) (Yang et al., 2018) is also a non-parameterized approach that requires zero training and zero parameters to construct sentences' representations. Inspired by the Gram-Schmidt Process in geometric theory, authors have developed an innovative method based on an orthogonal basis to combine pre-trained word embeddings into sentence representations. They have built an orthogonal basis of the subspace spanned by word embeddings and its surrounding context in the sentence to which it belongs. The semantic meaning of a word in a sentence is modeled based on two aspects: its relatedness to the word vector subspace already spanned by its contextual words, and the word's novel semantic meaning, which is introduced as a new basis vector perpendicular to the existing subspace.

2.2. Extractive Text Summarization

Traditional unsupervised extractive text summarization approaches are generally classified into statistical-based, graph-based, optimization-based, latent semantics-based and centroid-based approaches. The statistical-based approach combines statistical features to score sentences, where highly scored sentences are directly extracted to form the summary. These methods do not require any complex linguistic processing, and thereby, they are independent of the language and the domain of the text to summarize. We distinguish between sentence-level features such as sentence position, sentence length, sentence centrality, sentence resemblance to the title, etc. Word-level features attribute, though, a score for each word in the sentence, such as TF-IDF (Term-frequency-Inverse Document Frequency), mutual information, information gain, and residual inverse document frequency. Ferreira et al. (2013); Oliveira et al. (2016) have provided a comparative analysis of these features by applying them on single and multi-document summarization tasks. Graph-based approach (Erkan and Radev, 2004; Mihalcea and Tarau, 2004; Baralis et al., 2013) aims to represent document units (words, sentences) as a graph, and then apply graph-based ranking algorithms to generate the summary. Indeed, complex networks have shown to be useful for text summarization. For instance, Amancio et al. (2012) have developed an extractive summarizer that combines complex network analysis with language-dependent text processing. Where, they have employed new statistical metrics, namely betweenness, vulnerability, closeness, and diversity, to select relevant sentences. They have demonstrated that diversity measurement is more advantageous in generating informative summaries than the other centrality metrics. In the same context, Tohalino and Amancio (2018) have proposed a novel extractive method for multi-document summarization based on complex networks and multilayer networks. Where nodes represent sentences and edges are created by computing the similarity between two sentences using several measurements such as degree, strength, page rank, shortest path, and among others. Then, a multilayer network is established by considering each document as a layer. Unlike the previous works, they have demonstrated that such a distinction between intra-layer and inter-layer leads to better performance. Optimization-based approach (McDonald, 2007; Gillick et al., 2008; Metzler and Kanungo, 2008; Garcia et al., 2018) models the summarization task as an optimization problem using methods such as the integer linear programming (ILP), the sparse optimization, and the constraint optimization. While the Latent Semantic Analysis (LSA) based approach (Steinberger and Jezek, 2004) aims to analyze the semantic relationships between a cluster of documents and the terms it contains, it applies the Singular Value Decomposition (SVD) to extract the salient sentences of the documents.

The centroid-based approach aims to score sentences by computing the similarities between these sentences and the centroid. The centroid is defined as "a set of words that are statistically important to a cluster of documents" (Radev et al., 2004). Thus, the main idea of this approach is to build a pseudo_cluster that encodes the most meaningful information of a cluster of documents. The original centroid-based method for extractive multi-document summarization has been proposed by Radev et al. (2004). Where, each sentence S_i in a cluster of documents is represented as a Bag-of-Words (BOW) vector with the $tf - idf$ weighting scheme (Ramos et al., 2003), noticing that the dimension of vectors

is equal to the documents'vocabulary size. After, words with the $tf - idf$ greater than a topic threshold are selected as centroid keywords. Hence, the centroid vector is obtained by calculating the mean[2] of the keywords vectors that are considered central to all the articles in the cluster. Finally, the relevance of each sentence in the cluster is measured in relation to its cosine similarity to the centroid of the cluster, and then the top-ranked sentences are iteratively selected to form the summary taking into consideration the desired summary length. Indeed, this method is usually used as a baseline for evaluating extractive multi-document summarization. Other experiments have shown that applying some modifications to this baseline method can perform quite well (Ghalandari, 2017). For instance, Rossiello et al. (2017) have proposed to improve the original centroid-based method relying on three major modifications: (i) exploiting the compositional capabilities of word embedding representations instead of using BOW representations; (ii) focusing only on the most important terms of the input documents; and (iii) minimizing the redundancy by adopting the cosine similarity between sentences, where a new sentence is included in the summary only if it does not exceed a similarity threshold to any of the already included sentences. The obtained results have shown that this method performs better than the original centroid-based method. Moreover, Ghalandari (2017) has applied the centroid approach at the summary level rather than the sentence level. First, they select candidate sentences and represent the summary as the centroid of its sentences. Then, a greedy algorithm is used to maximize the similarity between the centroid of the summary and the centroid of the cluster of documents in order to find the best summary under a length constraint. The obtained results have also illustrated higher performance over the original centroid-based method.

Recently, extractive approaches based on supervised and unsupervised deep learning architectures have gained much attention, providing significant results. For instance, Zhong et al. (2015) have proposed an unsupervised extractive method for query oriented multi-document summarization where Deep Restricted Boltzman Machines have been adopted to learn hierarchical concept representations. Based on these concepts representations, sentences are scored and selected to form the summary. Yousefi-Azar and Hamey (2017) have also proposed an unsupervised extractive query-oriented summarization method for single-document that is based on a stochastic version of deep auto-encoders called the Ensemble Noise Auto-Encoders (ENAE). The ENAE adds some noise to the input text representation and then select the relevant sentences from an ensemble of noisy runs. Joshi et al. (2019) have developed SummCoder, an unsupervised framework for extractive generic single-document summarization based on deep auto-encoders, this method differs from the method proposed by (Yousefi-Azar and Hamey, 2017) as it uses sentence embedding models to represent input documents instead of using term-Frequency (tf) features.

Text representation is a fundamental process that affects the effectiveness of text summarization methods. Traditional ATS methods are rule-based, and most of them rely on hand-crafted features. Recently, word embeddings that are based on deep neural networks have shown significant progress in automatic text summarization task. The representation power of neural networks is related to their ability to learn high-level features across multiple layers and create accurate decision boundaries for the input instances. Recently, Mohd et al. (2020) have proposed a novel method for extractive text summarization based on the word embedding representations and the k-means clustering algorithm (Hartigan and Wong, 1979). First, they represent the input sentences using the Word2Vec model (Mikolov et al., 2013); then, they apply the k-means algorithm to form clusters that are semantically related. And finally, they use a novel ranking algorithm based on various statistical features(i.g. sentence length, sentence position, etc.) to select the relevant sentences. Furthermore, other ATS researchers have adopted methods for learning sentence representations instead of word representations because sentence representations are able to capture the semantic relationships among words. For this purpose, several works exploit Convolutional Neural Networks architectures. For instance, Denil et al. (2014) have proposed a hierarchical convolutional model to introspect the structure of the document. This model consists of a sentence-level component and a document-level component. At sentence-level, a ConvNet is used to learn sentence representation based on their words embeddings. At the document-level, another Convolutional Network is applied to learn the entire document representation based on their sentences' embeddings obtained in the first level. Then based on these representations, sentences are scored and selected to form the summary. Yin and Pei (2015) have also introduced a method based on convolutional neural networks where each sentence is projected to a continuous vector space, then an optimization process is run to select relevant sentences taking into consideration their "diversity" and "prestige" cost. Moreover, Cao et al. (2015) have developed a system based on enhanced convolutional neural networks, which automatically learn summary prior features for extractive summarization task.

Other automatic text summarization methods use pre-trained sentence embedding models that seek to map sentences into dense vectors that encode the semantics of these sentences. Kågebäck et al. (2014) have introduced a novel

[2]Other works use the sum to compute the centroid vector of a document or a cluster of documents. The similarity value does not change when using the cosine similarity since the angle between vectors remains the same/

method of sentence embedding representations based on submodular optimization (Lin and Bilmes, 2011) for the problem of multi-document summarization. Kobayashi et al. (2015b); Yogatama et al. (2015) have used pre-trained sentence embedding models for extracting relevant sentences following an unsupervised optimization paradigm. Furthermore, Cheng and Lapata (2016) have also proposed a data-driven approach based on neural networks and continuous sentence features for single document extractive text summarization task. Yasunaga et al. (2017) have introduced a neural multi-document summarization system; it is based on a Graph Convolutional Network (GCN) that takes sentences' embeddings obtained from Recurrent Neural Networks as input node features. Zhang et al. (2017) have proposed a sentence vector encoding framework based on a deep LSTM model; it embeds sentences into continuous vectors for single-line text summarization task.

In contrast to the existing centroid-based methods for extractive multi-document summarization (Radev et al., 2004; Rossiello et al., 2017; Ghalandari, 2017), we explore the potential of transfer learning from sentence embedding models to improve the performance of sentence representation and thus boost the performance of the proposed method. The idea is justified by the fact that transfer learning helps in saving time and computational power as well as benefiting from knowledge learned from other natural language understanding tasks. Moreover, we showcase how the combination of the sentence content relevance, sentence novelty, and sentence position metrics improves the performance of extractive multi-document summarization. Finally, we provide a detailed empirical study of the recent sentence embedding models on extractive multi-document summarization task that elucidates their interesting aspects and presents some differences between them.

3. Proposed Method

In this work, we propose an unsupervised method for extractive multi-document summarization based on the centroid approach and the sentence embedding representations. We use sentence embedding models to represent input documents and to build the centroid vector related to a cluster of documents. We adopt three essential metrics to compute sentences' scores and to determine which sentences are relevant for the summary. The proposed method consists of five main steps: (1) preprocessing, (2) sentence embedding, (3) centroid embedding, (4) sentence scoring, and (5) sentence selection. The overall architecture of the proposed method is illustrated in Figure 1 while its procedure is reported in Algorithm 1.

Algorithm 1 Pseudo code of the proposed method

Input: ***Sents***: an array of N sentences of the cluster D; ***sentPositions***: an array of sentences positions; ***docLens***: a dictionary of documents lengths; ***embeddingModel***: sentence embedding model; ***limit***: summary length; τ: sentence similarity threshold, α, β, and λ: weighting parameters
Output: summary
Begin
  finalScores, ***contentRelevanceScores***: N-dimensionanl arrays ▷ Local variable
  $sentVectors \leftarrow embeddingModel.embedSentences(Sents)$ ▷ Sentences embeddings
  $clusterCentroid \leftarrow mean(sentVectors)$ ▷ Computing the centroid vector of the cluster D
  for $i \leftarrow 1$ to N **do**
    $contentRelevanceScores[i] \leftarrow score^{contentRelevance}(sentVectors[i], clusterCentroid)$
  end for
  for $i \leftarrow 1$ to N **do**
    $noveltyScores \leftarrow score^{novely}(sentVectors, i, contentRelevanceScores, \tau)$
    $M^d \leftarrow docLens(i)$ ▷ The length of the document d containing sentence i
    $positionScores \leftarrow score^{position}(sentPositions[i], M^d)$
    $finalScores[i] \leftarrow \alpha * contentRelevanceScores[i] + \beta * noveltyScores + \lambda * positionScores$
  end for
  $summary \leftarrow generateSummary(Sents, finalScores, limit)$
  return $summary$
End

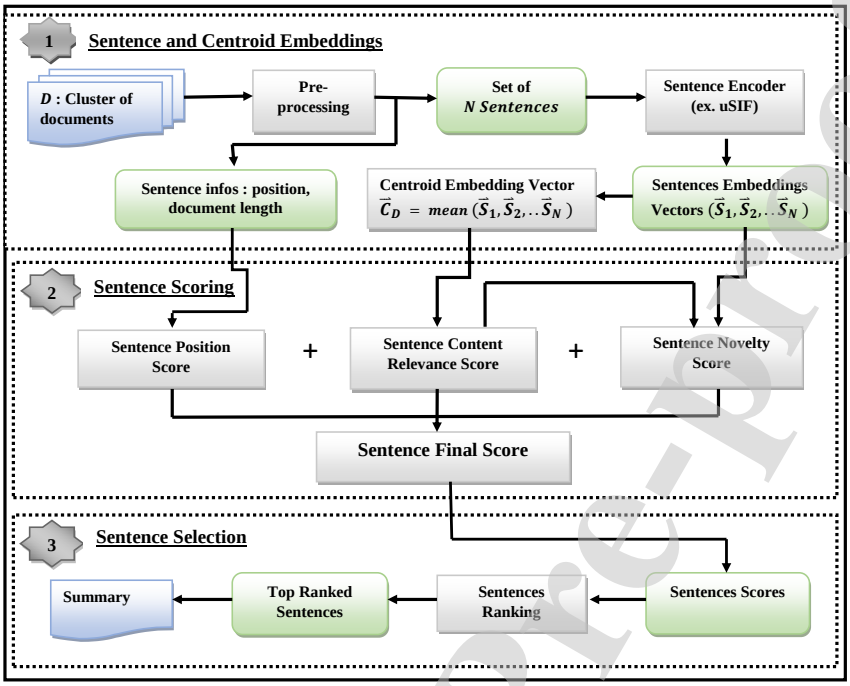

Figure 1: Architecture of the proposed method for extractive multi-document summarization.

3.1. Preprocessing

In the preprocessing step, we follow the commonly used pipeline for multi-document summarization task where we represent a cluster of documents by a set of sentences. Formally given a cluster D containing m documents $D = [d_1, d_2, ..., d_m]$. First, we split each document d_i from the cluster D into sentences using the open-source software library for Advanced Natural Language Processing spaCy[3]. Then, we use Natural Language Toolkit[4] (NLTK) and regular expressions to clean these sentences by converting all words in lower case as well as removing special characters, redundant whitespaces, XML/HTML tags, URLs and email addresses. And thus, we finally obtain a cluster D of N sentences, formally denoted as $D = [S_1, S_2, ..., S_N]$.

3.2. Sentence Embedding

After the preprocessing step, the next step is to map each sentence S_i in the cluster D into a fixed-length vector $\overrightarrow{S_i^D}$ using one of the pre-trained sentence embedding models described in section 2.1 (e.g. uSIF, DAN, BERT, ...). Generally, there are two strategies to use these models for Natural Language Processing (NLP) tasks: (i) *feature-based approach*, which uses the pre-trained models to extract fixed features in order to use them as input to NLP tasks. (ii) *Fine-tuning based approach*, which trains the downstream tasks by fine-tuning the pre-trained sentence embedding models parameters. Since we propose an unsupervised method for extractive multi-document summarization, we adopt the feature-based approach where the pre-trained model is applied on the cluster of the documents to be summarized. This shows the particular advantage of representation learning that allows us to reuse external knowledge, especially when we do not have a large amount of training data.

3.3. Centroid Embedding

In this step, we build the centroid embedding vector of the cluster D. Formally, given a cluster D of N sentences $D = [S_1, S_2, ..., S_N]$, where each sentence S_i in the cluster D is represented by an embedding vector $\overrightarrow{S_i^D}$ (see previous section). Thus, to construct the centroid embedding vector $\overrightarrow{C_D}$ of the cluster D, we compute the mean vector of this

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

cluster's sentences embeddings vectors, described as follows:

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

Where $\overrightarrow{C_D}$ denotes the centroid embedding related to 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 idea behind using sentence embedding representations to build the centroid relies on the fact that Bag-of-Words and word embeddings representations lose the word order as well as they ignore the sentence semantic by neglecting the semantic relationships between words in sentences. Figure 2 presents a conceptual representation of the centroid-based approach for extractive text summarization.

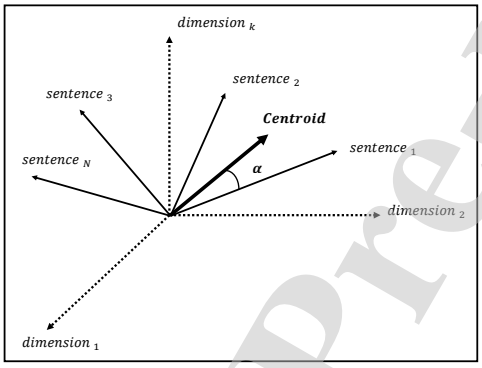

Figure 2: Conceptual representation of the centroid-based approach for extractive multi-document summarization. A cluster of documents D is represented as a set of N sentences. Each sentence S_i is represented in a space of k-dimensional vector space using a sentence embedding model. The centroid vector of the cluster D is computed as the mean of the cluster's sentences embeddings vectors.

3.4. Sentence scoring

Sentence scoring represents the fundamental cornerstone of extractive text summarization methods; it assigns a score for each sentence in the cluster, which helps to decide whether a sentence is relevant to the summary or not. In our case, we propose to score each sentence S_i in the cluster D by combining three metrics. (1) The sentence content relevance; (2) the sentence novelty; and (3) the sentence position. Formally, we consider a cluster of documents D composed of N sentences $D = [S_1, S_2, ..., S_N]$. Let's denote the sentence embedding vector of the sentence S_i in the cluster D as $\overrightarrow{S_i^D}$, and the centroid embedding vector of the cluster D as $\overrightarrow{C_D}$, where both the $\overrightarrow{S_i^D}$ and the $\overrightarrow{C_D}$ represent real-valued vectors in a k-dimensional Euclidian space R^k. Hence, in the following subsections, we will describe in detail the three metrics used to score each sentence S_i in the cluster D.

3.4.1. Sentence content relevance score

The proposed method makes use of the sentence embedding vector $\overrightarrow{S_i^D}$ and the centroid embedding vector $\overrightarrow{C_D}$ (See section 3.2 and section 3.3) to compute the sentence content relevance score. The centroid aims to condense the meaningful information of the cluster in one vector. Hence, it is reasonable to assume that the sentences' vectors, which are more similar to the centroid vector, are considered more relevant than the other sentences in the cluster. Thus, to obtain the sentence content relevance score of the sentence S_i in the cluster D, we compute the cosine similarity between the sentence embedding vector $\overrightarrow{S_i^D}$ and the centroid embedding vector $\overrightarrow{C_D}$ as described in equation 2.

$$score^{contentRelevance}(S_i, D) = cosineSimilarity(\overrightarrow{S_i^D}, \overrightarrow{C_D}) = \frac{\overrightarrow{S_i^D}.\overrightarrow{C_D}}{\| \overrightarrow{S_i^D} \| . \| \overrightarrow{C_D} \|} \tag{2}$$

Where $score^{contentRelevance}$ represents the content relevance score of the sentence S_i in the cluster D, $\overrightarrow{C_D}$ is the centroid embedding vector of the cluster D, and $\overrightarrow{S_i^D}$ is the embedding vector of the sentence S_i. The $score^{contentRelevance}$ is

bounded in [0,1], where sentences with higher scores are considered more relevant.

3.4.2. Sentence novelty score

The main purpose of using sentence novelty metric is to deal with redundancy and to produce summaries with good information diversity. In this work, we adopted the sentence novelty score proposed by Joshi et al. (2019). The sentence novelty metric computes the novelty of each sentence S_i in the cluster D; it assigns to the sentence a low score when it is redundant and a high score when it is novel. Thus, in order to get the novelty score of a sentence S_i in the cluster D. First, we compute its similarity with all the other sentences in the cluster D by measuring the cosine similarity between their corresponding embedding vectors, as described in equation 3. Where, $\overrightarrow{S_i^D}$ and $\overrightarrow{S_k^D}$ represent the embedding vectors of the sentence S_i and the sentence S_k respectively, and N is the number of sentences in the cluster D.

$$sim(S_i, S_k) = cosineSimilarity(\overrightarrow{S_i^D}, \overrightarrow{S_k^D}) = \frac{\overrightarrow{S_i^D}.\overrightarrow{S_k^D}}{\| \overrightarrow{S_i^D} \| . \| \overrightarrow{S_k^D} \|}, 1 \leq k \leq N, i \neq k \tag{3}$$

Then, if the maximum of the obtained similarities $sim(S_i, S_k), 1 \leq k \leq N, i \neq k$ is below a given threshold τ, then the sentence S_i is considered novel. When the similarity between two sentences of the cluster D is greater than the given threshold, the sentence with the higher content relevance score gets the higher novelty score. The sentence novelty score (Joshi et al., 2019) is calculated as follows:

$$score^{novelty}(S_i, D) = \begin{cases} 1, \text{if } \max(sim(S_i, S_k)) < \tau, 1 \leq k \leq N, i \neq k \\ 1, \text{if } \max(sim(S_i, S_k)) > \tau \text{ and } score^{contentRelevance}(S_i, D) > score^{contentRelevance}(S_l, D), \\ \qquad\qquad l = \underset{sim(S_i, S_k)}{\arg\max}\ (sim(S_i, S_k)),\ 1 \leq k \leq N,\ i \neq k \\ 1 - \max(sim(S_i, S_k)), otherwise, \end{cases} \tag{4}$$

Where $sim(S_i, S_k)$ reprsents the similariy between the sentence S_i and the other sentences in the cluster D, as described in equation 3. l is the argmax of the $sim(S_i, S_k)$, which means that l represents the index of the sentence that is the most similar to the sentence S_i in the cluster D. The $score^{novelty}$ and the $sim(S_i, S_k)$ are bounded in [0,1]. τ represents the threshold, in order to determine the best value of τ, we have tested several values of τ, namely the values comprised between $[0.5, 0.95]$ with constant steps of 0.05.

3.4.3. Sentence position score

Sentence position has been introduced first by Edmundson (1969); it is considered as one of the most effective sentence-based heuristics for selecting relevant sentences, especially for news articles (Oliveira et al., 2016). In our method, we use the sentence position scoring metric that is introduced by Joshi et al. (2019). Formally, given D a cluster of m documents, and each document d consists of M sentences. The sentence position score is computed, as shown in equation 5.

$$score^{position}(S_i^d) = \max(0.5, \exp(\frac{-p(S_i^d)}{\sqrt[3]{M^d}})) \tag{5}$$

Where, $score^{position}(S_i^d)$ represents the position score of a sentence S_i in a document d, $-p(S_i^d)$ is the ith position of S in d with $p(S_i^d)$ starting by 1, and M^d is the number of sentences in the document d. The obtained score is bounded in $[0.5, 1]$, assuming that the first sentences in a document are the most relevant. The importance of sentences decreases when the sentence goes far from the beginning of the document, noticing that the score remains constant at the value of 0.5 after a number of sentences.

3.5. Sentence selection

In order to generate summaries with good information diversity, we score each sentence S_i in the cluster D by combining linearly three scores including the sentence content relevance score (equation 2), the sentence novelty score

(equation 4), and the sentence position score (equation 5). Hence, we assume that relevant sentences are those that maximize the weighted sum of the three scores. Formally, the final score of a sentence S_i in the cluster D, denoted as $score^{final}(S_i, D)$, is defined by the following equation:

$$score^{final}(S_i, D) = \alpha * score^{contentRelevance}(S_i, D) + \beta * score^{novelty}(S_i, D) + \lambda * score^{position}(S_i^d) \quad (6)$$

Where, $\alpha+\beta+\lambda = 1$ with $\alpha, \beta, \lambda \in [0, 1]$ with constant steps of 0.1. Finally, the top ranked sentences are iteratively selected to form the final summary respecting the compression rate (pre-given summary length).

4. Experimental results

In this section, we present a comparative analysis of the results obtained by the proposed method according to the used sentence embedding models and the sentence salience scoring techniques. Hence, the experiments were performed to address the followings issues: (1) Evaluating the use of sentence embedding models on extractive multi-document summarization task; (2) Investigating the impact of the used sentence salience scoring techniques on the performance of sentence scoring phase; (3) Comparing the centroid-based method using three different representations (bag-of-words, word embeddings, and sentence embedding); and (4) Assessing the performance of the proposed method in contrast to the supervised and unsupervised state-of-the-art methods. Before presenting the experimental results, we provide a brief description of the used datasets, the evaluation measures, the experimental setup, and the approaches used for comparative analysis.

4.1. Datasets

In this work, we evaluate our method on four datasets for generic multi-document summarization including the standard DUC'2002-2004 datasets and the recently released Multi-News dataset (Fabbri et al., 2019). DUC (Document Understanding Conference) datasets are created by NIST[5] (National Institute of Standards and Technology) and considered as the widely used corpora for evaluating text summarization. DUC'2002 and DUC'2003 datasets contain 59 clusters and 30 clusters, respectively. Each cluster consists of approximately 10 English articles of news, distributed by TREC. While DUC'2004 Task 2 dataset includes 50 clusters, where each set consists of 10 documents, coming from Associated Press and New York Times newswires. For each cluster, four summaries written by different experts are provided. Table 2 describes the three underlying datasets. Besides, the Multi-News dataset represents the first large scale multi-document news summarization dataset. It contains about 44972 pairs for training, 5622 pairs for development, and 5622 for the test. The average of summaries is about 264 words paired where each summary is paired with a cluster of documents of average 2103 words.

Table 2
A description of DUC'2002, DUC'2003 and DUC'2004 datasets (Dernoncourt et al., 2018)

Dataset	Domain	Clusters	Documents	Sentences	Tasks
DUC'2002	News	59	576	14 370	Generic single and multi-document
DUC'2003	News	30	309	7691	Generic single and multi-document
DUC'2004	News	50	500	13135	Generic multi-document

4.2. Evaluation measures

For evaluation, we have adopted the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004), particularly ROUGE-N (ROUGE-1, ROUGE-2, and ROUGE-4), ROUGE-SU4, and ROUGE-L. ROUGE is a fully automated and state-of-the-art method for text summarization evaluation.
ROUGE-N measures the similarity between system summaries and a collection of summaries models (human summaries) based on the n-gram comparison and overlap. It is computed as follows:

$$ROUGE - N = \frac{\sum_{S \in (ReferenceSummary)} \sum_{N-gram \in (S)} Count_{match}(N - gram)}{\sum_{S \in (ReferenceSummary)} \sum_{N-gram \in (S)} Count(N - gram)} \quad (7)$$

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

Where, N is the length of $N-gram$, $Count_{match}(N-gram)$ is the maximum number of $N-grams$ that occur in both gold summary and candidate summary. ROUGE-1 and ROUGE-2 are the most used ROUGE measures, and they calculate the number of overlapping unigrams and bigrams respectively.
ROUGE-SU4 measures the overlap of skip-bigram between a system summary and a set of reference summaries with a max distance of four words. While ROUGE-L evaluates the fluency of the summary, it is based on the Longest Common Subsequence (LCS) that takes into account the sentence level structure similarity. Let us denote X a candidate summary and Y a gold summary that contains n words. ROUGE-L is calculated as follows:

$$ROUGE-L = \frac{LCS(X,Y)}{n} \tag{8}$$

Where, $LCS(X,Y)$ is the length of the longest subsequence of X and Y.
We have calculated ROUGES scores with the ROUGE toolkit (version 1.5.5), adopting the same ROUGE settings[6] that are used on DUC and Multi-News datasets for multi-document summarization. Following the state-of-the-art methods guidelines, we report Recall and F1-score values of ROUGE evaluation metrics on DUC datasets and the Multi-News dataset respectively.

4.3. Experimental setup

The proposed method has been developed using Python, relying on Tensorflow[7] and Keras[8] libraries. To generate sentences embeddings, we used the pre-trained models described in Section 2.1 on DUC'2002, DUC'2003, and DUC'2004. Where each model is used to embed a sentence into a dense vector; in other words, we extract the sentence embedding vectors and use these representations directly in our method, without any additional fine-tuning. Table 3 presents a concise comparison of the different sentence embedding models evaluated in this work.

Table 3
Comparison of the different sentence embedding models evaluated in this work. The embeddings size are the final sentence embedding after applying word embedding methods (e.g. word2vec)

Models	Learning pardigms	Training method	Word Vectors	Embedding size	Datasets
uSIF	Non-parameterized	Unsupervised	ParaNMT	300	– – –
GEM_GloVe	Non-parameterized	Unsupervised	GloVe	300	– – –
GEM_LexVec	Non-parameterized	Unsupervised	LexVec	300	– – –
ELMo	Parameterized	Supervised	– – –	1024	One Billion Word Benchmark
NNLM	Parameterized	Supervised	– – –	128	English Google news
USE-DAN	Parameterized	Both	Skip-gram	512	SNLI, Wikipedia, news, web pages ...
USE-Transformer	Parameterized	Both	– – –	512	SNLI, Wikipedia, news, web pages ...
InferSent_GloVE	Parameterized	Supervised	GloVe	4090	SNLI dataset
InferSent_FastText	Parameterized	Supervised	FastText	4090	SNLI dataset
$BERT_{BASE}$	Parameterized	Unsupervised	– – –	768	BookCorpus and Wikipedia
$BERT_{LARGE}$	Parameterized	Unsupervised	– – –	1024	BookCorpus and Wikipedia
Skip-Thought	Parameterized	Unsupervised	CBOW	2400	BookCorpus

The final score of a sentence is computed through a combination of three scoring methods: the sentence content relevance, the sentence novelty, and the sentence position scores, as described in equation 6. Where, α, β, and λ are comprised between $[0,1]$ with constant steps of 0.1 and the threshold τ is comprised between $[0.5, 0.95]$ with constant step of 0.05. In order to determine the values of these hyperparameters, we employed a similar procedure to the one that is used by Joshi et al. (2019). We built a small held-out set by shuffling and randomly sampling 25 clusters from the validation set of the Multi-News 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. Finally, we generated the summaries of DUC'2002-2004 and the Multi-News datasets using the latter combination. Moreover, we performed the paired

[6] ROUGE-1.5.5 with parameters "-n 4 -m -l 100 -c 95 -r 1000 -f A -p 0.5 -t 0" (DUC), "-a -c 95 -m -n 2 -2 4 -u -p 0.5 -l 300 "(Multi-News)
[7] https://www.tensorflow.org/
[8] https://keras.io/

Student's t-test for statistical significance testing and attached a superscript to the performance number in the tables when the $p-$ value < 0.05.

4.4. Proposed method evaluation

Firstly, the experiments are conducted to investigate the performance of the three sentence scoring measures as well as to assess the impact of their combination. The main goal of these experiments is to answer the following research question: 1) *Does the combination of the sentence novelty metric, the sentence position metric, and the centroid-based method improve the performance?*

To answer the latter question, we have performed four runs on DUC'2004 dataset, described as follows:

- Run 1 : the sentence content relevance score (centroid based method);
- Run 2: the combination of sentence content relevance and position scores;
- Run 3: the combination of sentence content relevance and novelty scores;
- Run 4: the combination of the three sentence scoring measures.

Table 4 summarizes the obtained results of the four runs. Moreover, a concrete example of the system summaries generated by these four runs using uSIF model along with the reference summaries is illustrated in Table 8 in Section A.

Table 4
The obtained results of the different scoring measures and their combination using nine different sentence embedding models on the extractive multi-document summarization dataset DUC'2004. For denoting statistical significance results, the superscripts 1, 2, 3, and 4 indicate significant improvement ($p-$ value < 0.05) over Run 1, Run 2, Run 3, and Run 4, respectively.

Models	Run 1			Run 2			Run 3			Run 4		
	R-1	R-2	R-4	R-1	R-2	R-4	R-1	R-2	R-4	R-1	R-2	R-4
uSIF	39.03	9,06	1.39	39.67	9.48	1.54	39.21	9.35	1.61	39.72	$9.79^{1,3}$	1.65^{1}
USE-DAN	39.45	8.18	1.12	$39.94^{1,3}$	$9.75^{1,3}$	$1.51^{1,3}$	39.48	8.76	1.18	$40.14^{1,3}$	$9.85^{1,3}$	$1.58^{1,3}$
USE-Transformer	37.49	7.93	0.87	$39.47^{1,3}$	$9.48^{1,3}$	$1.42^{1,3}$	38.93	8.06	1.12	$39.84^{1,3}$	$9.53^{1,3}$	$1.48^{1,3}$
NNLM	37.05	7.93	0.97	$39.19^{1,3}$	$9.02^{1,3}$	$1.35^{1,3}$	38.73^{1}	8.96^{1}	1.14^{1}	$39.45^{1,3}$	$9.19^{1,3}$	$1.44^{1,3}$
ELMo	36.83	7.86	0.89	38.23^{1}	$8.69^{1,3}$	1.21^{1}	37.64	8.08	1.09	38.35^{1}	$8.76^{1,3}$	1.25^{1}
InferSent_GloVe	38.43	8.53	1.24	38.61	8.77^{1}	1.32	38.55	8.58^{1}	1.26	38.71^{1}	$9.17^{1,3}$	1.38^{1}
InferSent_FastText	33.43	5.64	0.63	$34.39^{1,3}$	$6.64^{1,3}$	0.93	33.62	5.95	0.76	$34.89^{1,2,3}$	$7.08^{1,2,3}$	1.01^{1}
Skip-Thought_Unid	36.09	7.49	0.73	36.68	8.04^{1}	1.08	36.19	7.97	0.98	36.73^{1}	8.06^{1}	1.12^{1}
Skip-Thought_Bid	37.29	8.36	0.94	37.53	8.65	1.25	37.47	8.43	1.09	37.54^{1}	$8.72^{1,3}$	1.32^{1}
GEM_LexVec	33.73	5.45	0.53	34.06	6.14	0.62	33.95	5.52	0.59	34.08^{1}	6.38^{1}	0.77^{1}
GEM_GloVe	33.75	5.89	0.71	34.28	6.87	0.97	34.18	6.58	0.83	34.82^{1}	6.97^{1}	0.98^{1}
BERT_{BASE}	27.76	3.66	0.23	28.82^{1}	4.34^{1}	0.56^{1}	27.8	4.18	0.31	28.92^{1}	4.43^{1}	0.59^{1}
BERT_{LARGE}	23.17	3.25	0.26	24.85^{1}	3.67^{1}	0.32^{1}	24.38	3.31	0.28	24.89^{1}	3.83^{1}	0.34^{1}

Table 4 shows clearly that the content relevance measure achieves a good performance, especially for the uSIF, USE-DAN, USE-Transformer, NNLM, InferSent_GloVe, and the bidirectional Skip-Thought (Skip-Thought_Bid) models in terms of R-1, R-2, and R-4 scores. Although the combination of the content relevance and the novelty scores (Run 3) improved the performance, the overall obtained results do not show significant improvement over the sentence content relevance scoring method (Run 1). Moreover, the combination of the content relevance and sentence position scores (Run 2) has lead to significant improvements over the sentence content relevance (Run 1) and its combination with sentence novelty (Run 3) for the majority of the used embedding models. The latter can be explained by the fact

that the first sentences of news articles constitute the most relevant and vital parts of these articles' content. Finally, the combination of the three sentence scoring measures (Run 4) has provided a significant improvement for most of the used sentence embedding models. Thus, these three metrics are complementary to each other and improve the performance of the multi-document summarization task. Indeed, incorporating sentence novelty and position scores to the centroid-based method significantly improves the performance.

Secondly, we have performed several experiments to evaluate the performance of each sentence embedding model used in our method. These experiments aim to address the following research question: 2) *Which are the best sentence embedding models for the extractive multi-document summarization task?*

Table 5 presents the comparison results of the different sentence embedding models exploited in the proposed method, using the combination of the three sentence scoring measures on the three multi-document summarization datasets DUC'2002-2004. Moreover, Figure 3 shows the performance of the used sentence embedding models in terms of ROUGE-1, ROUGE-2, and ROUGE-4 recall scores.

Table 5
Comparison results of the different sentence embedding models using the combination of the three sentence scoring measures on the three extractive multi-document summarization datasets. For denoting statistical significance results, the superscripts *number* indicates significant improvement ($p - \text{value} < 0.05$) over the sentence embedding model that has the same superscript *number* attached. The interval $i - j$ indicate a significant improvement over models that have a superscript *number* attached ranging from i to j.

Models	DUC'2004			DUC'2003			DUC'2002		
	R-1	R-2	R-4	R-1	R-2	R-4	R-1	R-2	R-4
uSIF[1]	$39.72^{5,7-13}$	9.79^{4-13}	1.65^{5-13}	$38.29^{5,7,9-13}$	$9.27^{5,7-13}$	$1.48^{5,7-13}$	$37.42^{5,7-13}$	$8.19^{5,7-13}$	$1.29^{5,7-13}$
USE-DAN[2]	40.14^{5-13}	9.85^{4-13}	1.58^{5-13}	$38.35^{5,7,9-13}$	$9.06^{5,7-13}$	$1.28^{5,7-13}$	$37.56^{5,7-13}$	$8.24^{5,7-13}$	$1.18^{5,7-13}$
USE-Transformer[3]	$39.84^{5,7-13}$	$9.53^{5,7-13}$	1.48^{5-13}	$38.05^{5,7,9-13}$	$8.63^{5,7-13}$	$1.12^{5,7-13}$	$36.82^{5,7-13}$	$8.09^{5,7-13}$	1.12^{7-13}
NNLM[4]	$39.45^{5,7-13}$	9.19^{7-13}	$1.44^{5,7-13}$	$38.11^{5,7,9-13}$	$9.17^{5,7-13}$	$1.56^{5,7-13}$	$37.11^{5,7-13}$	$7.65^{7-8,10-13}$	$1.04^{7,10-13}$
ELMo[5]	38.35^{7-13}	$8.76^{7-8,10-13}$	$1.25^{7-8,10-13}$	$34.02^{7,12-13}$	6.23^{12-13}	0.74^{12-13}	$35.74^{7,10-13}$	$7.49^{7-8,10-13}$	$0.94^{7,10-13}$
InferSent_GloVe[6]	38.71^{7-13}	9.17^{7-13}	$1.38^{7-8,10-13}$	$37.59^{5,7,9-13}$	$9.01^{5,7-13}$	$1.47^{5,7-13}$	$37.16^{5,7-13}$	$8.07^{5,7-13}$	$1.19^{5,7-8,10-13}$
InferSent_FastText[7]	$34.89^{10,12-13}$	$7.08^{10,12-13}$	$1.01^{10,12-13}$	32.83^{12-13}	6.61^{12-13}	0.71^{12-13}	32.82^{12-13}	5.44^{12-13}	0.67^{12-13}
Skip-Thought_Unid[8]	$36.73^{7,10-13}$	$8.06^{7,10-13}$	$1.12^{10,12-13}$	$36.34^{5,7,10-13}$	$7.33^{5,7,10-13}$	0.81^{11-13}	$35.23^{7,10-13}$	$6.35^{7,10-13}$	0.71^{10-13}
Skip-Thought_Bid[9]	$37.54^{7,10-13}$	$8.72^{7-8,10-13}$	$1.32^{7-8,10-13}$	$35.83^{5,7,10-13}$	$7.13^{5,7,10-13}$	0.72^{12-13}	$36.27^{7,10-13}$	$7.23^{7,10-13}$	$1.06^{7,10-13}$
GEM_LexVec[10]	34.08^{12-13}	6.38^{12-13}	0.77^{12-13}	$34.18^{7,12-13}$	6.57^{12-13}	0.74^{12-13}	32.38^{12-13}	5.04^{12-13}	0.42^{12-13}
GEM_GloVe[11]	$34.82^{10,12-13}$	$6.97^{10,12-13}$	$0.98^{10,12-13}$	33.16^{12-13}	6.74^{12-13}	0.68^{12-13}	32.84^{12-13}	5.59^{12-13}	0.54^{12-13}
BERT^{12}_{BASE}	28.92^{13}	4.43^{13}	0.59^{13}	28.03^{13}	4.48^{13}	0.45^{13}	28.72^{13}	3.98^{13}	0.23^{13}
BERT^{13}_{LARGE}	24.89	3.83	0.34	21.74	2.89	0.24	22.96	3.12	0.21

For DUC'2002 dataset, the results show that uSIF, USE-DAN, USE-Transformer, NNLM, and InferSent_GloVE embedding models achieve the best performances and significantly outperform most of the other sentence embedding models for all evaluation measures (R-1, R-2, and R-4). Moreover, ELMo and Skip-Thought have also achieved good results and lead to significant improvements over the GEM_GloVe, GEM_LexVec, BERT_{BASE}, and BERT_{LARGE} models. Even though the bidirectional Skip-Thought outperformed its unidirectional version (Skip-Thought_Unid), the difference between the two models is not statistically significant. Furthermore, models that use GloVe word embedding as input to the sentence embedding model (InferSent_GloVe and GEM_GloVe) showed a better performance than their version that uses FastText and LexVec (InferSent_FastText and GEM_LexVec).

For DUC'2003 corpus, the obtained results show that uSIF, USE-DAN, USE-Transformer, InferSent_GloVe, and NNLM models have achieved the best performances and lead to significant improvements over most other sentence embedding models. Even though they outperformed the unidirectional Skip-Thought (Skip-Thought_Unid) models, the differences between them are not statistically significant for the R-1 evaluation measure. Moreover, the bidirectional skip-Thought (Skip-Thought_Bid) models has achieved significant improvements over ELMo, GEM_GloVe, GEM_LexVec, InferSent_FastText, BERT_{BASE} and BERT_{LARGE} models.

For DUC'2004 dataset, and based on R-1 and R-2 sores, the obtained results show that USE-DAN model has achieved the best performance and leads to significant improvements over most sentence embedding models except uSIF, USE-Transformer, and NNLM. Regarding the evaluation measure R-4, uSIF model significantly outperforms the other sentence embedding models, except USE-DAN, USE-Transformer, and NNLM models. Moreover, ELMo and InferSent_GloVe have also achieved good results and obtain significant improvements over InferSent_FastText, Skip-Thought, GEM_GloVe, GEM_LexVec, BERT$_{BASE}$ and BERT$_{LARGE}$ models. In accordance with the overall obtained results using DUC'2002 dataset, the bidirectional Skip-Thought outperformed its unidirectional version (Skip-Thought_Unid), but the difference between both of them is not statistically significant regarding R-1. Furthermore, models that use GloVe word embedding as input to the sentence embedding model (InferSent_GloVe and GEM_GloVe) yield to a better performance than their version that uses FastText and LexVec (InferSent_FastText and GEM_LexVec).

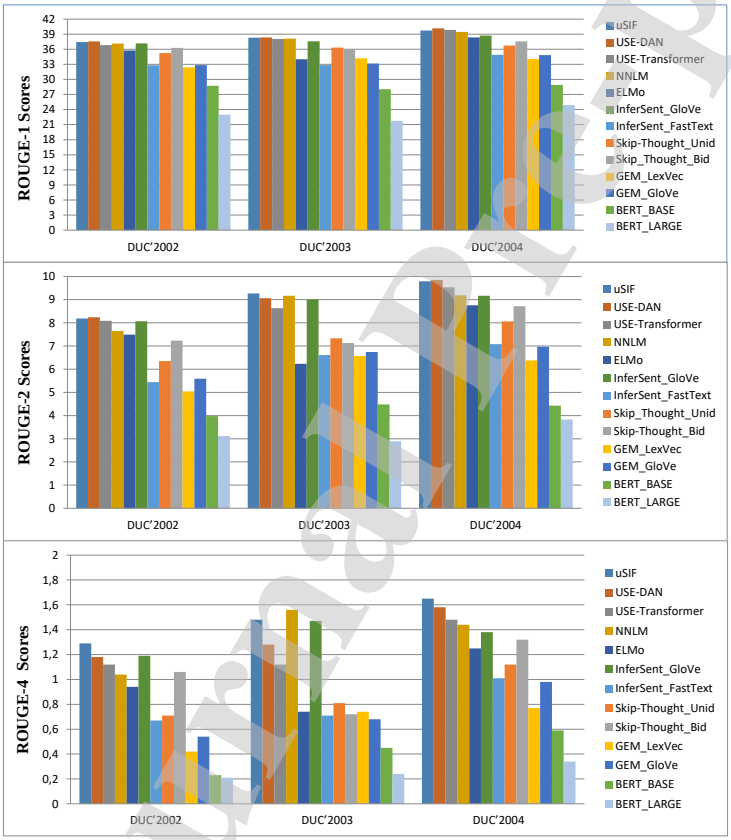

Figure 3: Examples of ROUGE-1, ROUGE-2 and ROUGE-4 scores obtained by our method according to the used sentence embedding models on DUC'2002, DUC'2003, and DUC'2004 .

The overall obtained results on the three datasets have demonstrated that uSIF, USE-DAN, USE-Transformer, NNLM, and InferSent_GloVe are the best and the most suitable sentence embedding models for unsupervised extractive multi-document summarization method. Although ELMo and Skip-Thought have achieved good performances, uSIF, USE-DAN, USE-Transformer and NNLM models significantly outperformed both of them on most datasets. Meanwhile, the comparison results show clearly that the performance of each sentence embedding model does not depend on the used text summarization dataset (DUC'2002, DUC'2003, and DUC'2004). This finding may be due to the fact that most used sentence embedding models are pre-trained using news corpus, as well as other text corpora like Wikipedia, whereas the documents of DUC datasets are news articles. Besides, some of these embedding models use word embeddings that are trained on news corpora as input for their pre-training. Indeed, the best performing models are the same for the three datasets. Furthermore, the results of Table 4 and Table 5 showed that the use of a suitable sentence embedding model, as well as the combination of the three sentence scoring measures (content relevance,

novelty, and position scores), have lead to significant improvements for all evaluation measures (R-1, R-2, and R-4).

4.5. Comparative evaluation with state-of-the-art methods

To evaluate the effectiveness of our method, we compare its performance with existing state-of-the-art methods for generic multi-document summarization using DUC'2004 and Multi-News datasets. Table 6 and Table 7 summarize the obtained ROUGE scores (R-1, R-2, R-4, R-SU4, and R-L) for each dataset.

First, we compare our method with three centroid-based methods for unsupervised extractive multi-document summarization. The original **Centroid_BOW method** (Radev et al., 2004) that uses bag-of-words representation. The **Centroid_WordEmbedding_v1 method** (Rossiello et al., 2017) that exploits the compositional capabilities of word embedding. And, the **Centroid_WordEmbedding_v2 method** (Ghalandari, 2017) the applies the centroid approach at the summary level to select the relevant sentences.

Then, we compare it with several extractive unsupervised methods from SumRepo (Hong et al., 2014), which is a repository of generic multi-document summaries generated by different systems for DUC'2004 dataset. The generated summaries are published in GitHub[9] while the code source of most of these systems is available in Sumy repository[10]. For instance, **Lead-3** is an extractive system that selects the first-3 sentences of each source document to form the final summary. **TextRank** (Mihalcea and Tarau, 2004) is an unsupervised graph-based ranking model that computes sentence importance scores for extractive text summarization. **LexRank** (Erkan and Radev, 2004) is also an unsupervised graph-based method for ATS that estimates the importance of sentences in the documents based on their centrality. **CLASSY04 [Peer65]** (Conroy et al., 2004) is based on Hidden Markov Models and the topic signature feature. It outperformed all the systems during the official DUC'2004 evaluation campaign. **ICSISumm** (Gillick et al., 2008) uses a global linear optimization framework to find the optimal summaries rather than greedily selecting relevant sentences. **DPPs summarization** (Kulesza et al., 2012) is based on probabilistic models of sets, namely the Determinantal Point Processes DPPs models. **ConceptBased_ILP** (de Oliveira et al., 2018) is based on the Integer Linear Programming to select relevant sentences where it exploits the combination of the centrality and the position metrics to extract salient concepts.

Finally, we compare it with recent supervised state-of-the-art deep learning based methods. **PG-MMR** (Lebanoff et al., 2018) is an adaptive method for multi-document summarization, which combines the pointer-generator network with the maximal marginal relevance criterion. **CopyTransformer** (Gehrmann et al., 2018) is based on Transformer and copying mechanism. Where a content selector is used to determine the sentences that should be in the summary, and a copying mechanism is used to pre-select sentences during decoding. **Hi-MAP** (Fabbri et al., 2019) is an end to end model for multi-document summarization, which incorporates maximal marginal relevance criterion into a hierarchical pointer-generator network. **DynE** (Hokamp et al., 2020) is a dynamic ensembling decoding method for multi-document summarization; it allows models trained on single inputs to be applied in multi-input settings. **MGSum** (Jin et al., 2020) applies a multi-granularity interaction network for multi-document summarization where it jointly learns semantic representations for words/sentences/documents.

The first set of analysis aims to compare the centroid-based method using three different representations (bag-of-words, word embeddings, and sentence embeddings). The main goal is to answer the following question: *Does the use of sentence embedding representations improve the performance of the centroid-based method?* The results of the experiments are shown in Table 6. We report the results of our method using the four best performing sentence embedding models (uSIF, USE-DAN, USE-Transformer, and NNLM). For all the sentence embedding models, our method significantly outperforms the original centroid-based method (Centroid_BOW). In detail, with the USE-DAN model, we obtain an improvement of 3,73%, 1,88%, 0,37%, and 3,74% with respect to the Centroid_BOW for R-1, R-2, R-4, and R-L metrics respectively. Additionally, for R-1 measure, our method achieves an increment of 2,23%, 1,81%, 1,93%, and 1,54% with respect to the Centroid_WordEmbedding_v1 method using USE-DAN, uSIF, USE-Transformer, and NNLM respectively. For R-2 and R-4 measures, our method shows comparable results to the Centroid_WordEmbedding_v1 method. Moreover, for the Centroid_WordEmbedding_v2 method, it yields an improvement of 1,03% regarding R-1 measure, while it shows comparable performances in terms of R-2 and R-4 evaluation measures. Therefore, the overall obtained results show the effectiveness of sentence embedding representations in improving the centroid-based method, where the aim is to map variable-length sentences to fixed-length vectors assuming that sentences with similar meanings have similar vectors, and simultaneously, sentences with different meanings have different vectors.

[9] https://github.com/stuartmackie/duc-2004-rouge
[10] https://github.com/miso-belica/sumy

Table 6
Systems performance on DUC'2004, using state-of-the-art methods and the proposed method. The highest performance for each of the group of methods is printed in boldface. The best performing method for each measure is indicated by *. For denoting statistical significance results, the superscripts *number* indicates significant improvement ($p-$value < 0.05) over the sentence embedding model 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. The results of models with ‡ attached are taken from their original articles.

Method	R-1	R-2	R-4	R-L
LexRank[1]	35.95	7.47	0.82	31.1
CLASSY 04[2]	37.62	8.96	1.51	32.26
ISCISum[3]	38.41	9.78	**1.73***	33.62
DPP[4]	**39.79**	9.62	1.57	**34.93**
ConceptBased_ILP[5]	38.65	**10.02***	1.67	34.23
Centroid_BOW[6]	36.41	7.97	1.21	31.21
Centroid_WordEmbedding_v1‡	37.91	9.53	1.56	–
Centroid_WordEmbedding_v2‡	**39.11**	**9.81**	**1.58**	–
PG-MMR‡	**36.42**	**9.36**	–	–
CopyTansformer‡	28.54	6.38	–	–
Hi-MAP‡	35.78	8.90	–	–
Proposed Method				
uSIF	$39.72^{1-3,5-6}$	$9.79^{1-2,6}$	$\mathbf{1.65}^{1,6}$	$\mathbf{34.98}^{*1-3,6}$
USE_DAN	$\mathbf{40.14}^{*1-3,5-6}$	$\mathbf{9.85}^{1-2,,6}$	$1.58^{1,6}$	$34.95^{1-3,6}$
USE_Transformer	$39.84^{1-3,5-6}$	$9.53^{1,6}$	$1.48^{1,6}$	$34.89^{1-3,6}$
NNLM	$39.45^{1-3,6}$	$9.19^{1,6}$	$1.44^{1,6}$	$34.69^{1-3,6}$

Table 7
ROUGE-F1 evaluation results on the Multi-News Dataset for the proposed method and the state-of-the-art methods. The results of the state-of-the-art methods are directly taken from their original articles. The highest performance for each of the group of methods is printed in boldface. The best performing method for each measure is indicated by *.

Method	R-1	R-2	R-SU4	R-L
Lead-3	**39.44**	11.77	**14.51**	–
LexRank	38.27	12.70	13.20	–
TextRank	38.44	**13.10**	13.50	–
PG-MMR	40.55	**16.36***	15.87	–
CopyTansformer	43.57	14.03	17.37	19.5
Hi-MAP	43.47	14.83	17.41	19.7
DynE	43.9	15.8	–	**22.2**
MGSum	**44.75***	15.75	**19.30***	–
Proposed Method				
uSIF	42.19	13.87	17.55	27.67
USE_ DAN	**42.93**	14.04	17.27	27.7
USE_Transformer	42.86	14.25	17.32	**28.48***
NNLM	42.62	**14.19**	**17.58**	27.93

The second set of analysis is conducted to compare our method with other extractive unsupervised multi-document summarization methods. The first block of Table 6 reports ROUGE scores (R-1, R-2, R-4, and R-L) of LextRank, CLASSY 04, ISCISum, DPP, and ConceptBased_ILP methods on DUC'2004 dataset. While the first block of Table 7 reports ROUGE scores (R-1, R-2, and R-SU4) of Lead-3, LexRank, and TextRank methods on the Multi-News dataset. As shown in Table 6, for R-1 measure, the variants of our method that use uSIF, USE-DAN, USE-Transformer have

significantly outperformed LextRank, CLASSY 04, ISCISum, and ConceptBased_ILP methods while it has achieved comparable result with DPP method, which is considered as the best performing method on DUC'2004. Moreover, our method with NNLM has significantly outperformed LextRank, CLASSY 04, and ISCISum methods, and achieves a comparable performance with ConceptBased_ILP and DPP methods. For R-2 measure, our method has achieved significant improvements over LexRank and CLASSY 04 methods and has obtained comparable results with ISCISum, DPP, and ConceptBased_ILP methods. Additionally, for the R-4 measure, our method has significantly outperformed LexRank and has achieved comparable results with the other methods. Finally, in terms of R-L measure, all variants of our method have achieved promising results; it has lead to significant improvements over LexRank, CLASSY 04, and ISCISum methods. Therefore, the overall results on the DUC'2004 dataset show that USE-DAN, ConceptBased_ILP, ISCISum, and uSIF are the best performing methods in terms of R-1, R-2, R-4, and R-L respectively. Furthermore, the obtained results, reported in Table 7, show that the variants of our method (uSIF, USE-DAN, USE-Transformer, and NNLM) have shown far better performances than Lead-3, LexRank, and TextRank methods on the Multi-News dataset. For instance, our method with USE-DAN has achieved an increment of 4,39%, 4,17%, and 2,76% with respect to the Lead-3 system in terms of R-1, R-2, and R-SU4 measures respectively.

The final set of analysis is performed to compare our method with the recent supervised deep learning based methods. The third block of Table 6 presents the obtained R-1 and R-2 scores of PG-MMR, CopyTransformer, and HI-MAP methods on DUC'2004 dataset. The second block of Table 7 reports ROUGE scores of PG-MMR, CopyTransformer, HI-MAP, DynE, and MGSum methods on the Multi-News dataset. As shown in Table 6, all variants of our method have outperformed PG-MMR, CopyTransformer, and HI-MAP methods. For instance, our method with USE-DAN has achieved an increment of 3,72% and 0,49% with respect to PG-MMR using R-1 and R-2 respectively. The obtained results can be explained by the fact that the supervised PG-MMR, CopyTransformer, and HI-MAP methods have been trained on CNN/DailyMail dataset, which is designed for single-document summarization, where documents are very short compared to a cluster of documents.
Meanwhile, Table 7 compares our method with supervised deep learning based methods (PG-MMR, CopyTransformer, HI-MAP, DynE, and MGSum) that are trained and tested on Multi-News dataset. In terms of R-1 and R-SU4 scores, our method outperformed the PG-MMR model, while the latter surpassed our method as well as all other state-of-the-art methods in terms of R-2. Besides, It has obtained comparable R-2 and R-SU4 performances with CopyTransformer and HI-MAP methods. For the R-1 measure, CopyTransformer, HI-MAP, DynE, and MGSum achieved better performance than our method, especially the MGSum model which yielded the best R-1 score. Finally, in terms of R-L measure, our method has achieved far better performances than CopyTransformer, HI-MAP, and DynE methods. The overall results show that the best performing method in terms of R-1 and R-SU4 is the MGSum model, while the PG-MMR yielded the best R-2 performance and our method which based on USE-Transformer achieved the best R-L score.

5. Discussion

The results presented in Table 4 demonstrate that the content relevance score (Run 1) has achieved a good performance. The latter might be explained by the effectiveness of the centroid-based method in condensing the meaningful information of the entire cluster, as well as the central importance of sentence embedding representations. Moreover, the fact that the combination of the content relevance and novelty scores (Run 3) has shown a better performance can be explained by the efficiency of the novelty metric in dealing with the redundancy issue. Since we address the multi-document summarization task, redundancy represents a critical problem. The risk to select sentences that convey the same information is more significant in comparison to the single-document text summarization task. Furthermore, the combination of the content relevance and position scores (Run 2) has shown significant improvements over the sentence content relevance score (Run 1) and its combination with the novelty score (Run3). These improvements are due to the use of the sentence position metric that is considered as one of the effective methods for selecting relevant sentences, especially in newswire documents. Indeed, it is known in the literature that for news articles, the most relevant information takes place at the leading sentences of the documents. The overall obtained results also illustrate that the combination of the three sentence scoring methods (Run 4) has significantly improved the performance of the proposed method for all the evaluated sentence embedding models. Thus, these three sentence scoring metrics are complementary to each other.

Another main objective of this work is to evaluate several pre-trained sentence embedding models on the extractive text summarization task. Indeed, most of these models have already been evaluated on the General Language

Understanding Evaluation (GLUE) benchmark (Wang et al., 2018) and applied to several NLP tasks. However, to the best of our knowledge, this is the first study that provides an empirical analysis of several sentence embedding models on the extractive multi-document text summarization task. The overall obtained results, illustrated in Table 5, show that the best performing models on extractive multi-document summarization are uSIF, USE-DAN, USE-Transformer, NNLM, and the InferSent_GloVe on the three DUC'2002-2004 datasets.

On the one hand, both USE-DAN and USE-Transformer encoders have been trained using a large amount of unsupervised data drawn from a variety of web sources documents, and for boosting the performance, the unsupervised learning is augmented with supervised training on the Stanford Natural Language Inference (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. Although USE-Transformer usually achieves better performance that USE-DAN in transfer learning tasks (Cer et al., 2018), USE-DAN performs better than the former in this work. The latter can be explained by the fact that the USE-DAN makes use of both sentence and word level transfer (Cer et al., 2018). Furthermore, the unsupervised model uSIF that uses ParaNMT word vectors outperformed many sophisticated neural sentence embedding models, including the USE-Transformer and the InferSent. Since uSIF has been tested on the SemEval Semantic Textual Similariy (STS) task and has achieved the state-of-the-art models on this task (Ethayarajh, 2018). The goal of this task lies in computing the semantic similarity between a given 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 to determine both the content relevance and the novelty scores. More interestingly, the probabilistic model NNLM has achieved promising results; it is based on deep neural networks trained using a large corpus of supervised data from the Associated Press (AP) News from 1995 and 1996. This model is considered as a strong sentence embedding baseline, which simultaneously learns words representations and a probability distribution for word sequences.

On the other hand, the unsupervised Skip-Thought has also achieved good results. These findings confirm that the Skip-Thought model learns representations that are well suited for semantic relatedness (Kiros et al., 2015). Moreover, the bidirectional Skip-Thought (Skip-Thought_Bid) model performs better than the unidirectional Skip-Thought (Skip-Thought_Unid) model. Indeed, the bidirectional Skip-Thought encoder reads the input sentence forward and backward, then concatenates the results (output vector) that are used by two decoders to predict the previous and the next sentence. Hence, the bidirectional Skip-Thought uses the full context of words of sentences. Although BERT has shown impressive results on many NLP tasks, its application to summarization using the feature-based approach is not straightforward. This finding may be due to the fact that BERT is trained on a masked-language model and require fine-tuning on the downstream task or other Natural Language Understanding task, like SNLI (used by USE-DAN and USE-Transformer to augment unsupervised training using supervised data, as well as a training data for InferSent). Moreover, the obtained results by two versions of the Geometric Embedding model (GEM_GloVe, GEM_LexVec) have also illustrated that this model does not perform well on multi-document summarization. Exploiting the geometric structure of the subspace spanned by word embeddings might not be beneficial for the underlying task. Furthermore, models that use GloVe word embedding as input to embedding model (InferSent_GloVe and GEM_GloVe) showed a better performance than their version that uses FastText and LexVec (InferSent_FastText and GEM_LexVec).

To summarize, sentence embedding models that are either trained using supervised data from the SNLI dataset or used the latter dataset to augment unsupervised training have been shown to be more effective for multi-document summarization. Therefore, the difference between the results comes from the different abilities and architectures of these models. Moreover, most sentence embedding models that achieve state-of-the-art results on the GLUE benchmark have shown good performance in this work.

Finally, the overall comparison results show that our method has achieved comparable performances with the best performing state-of-the-art methods (DPP, ISCISum and ConceptBased_ILP) on DUC'2004 dataset. Moreover, it has shown promising results on the Multi-News dataset in comparison to recent state-of-the-art deep learning based methods. Hence, these findings prove the effectiveness of sentence embedding models in capturing the semantic and the syntactic structure of sentences.

6. Conclusion

In this paper, we proposed an unsupervised extractive method for multi-document summarization based on the sentence embedding representations and the centroid-based approach. It consists of five main steps: (1) Preprocessing, (2) sentence embedding, (3) centroid embedding, (4) sentence scoring, and (5) sentence selection. In the first step, we

preprocess the input documents to represent a cluster of documents as a set of sentences. In the second step, we map each sentence of this cluster into a dense vector using one of the sentence embedding models exploited in this work. Then, we build the centroid embedding vector of this cluster by computing the mean vector of the cluster's sentences' embeddings vectors. In the next step, we combine the content relevance, novelty, and position measures to assign a score to each sentence in the cluster. Finally, we select the top-ranked sentences iteratively to form the summary.

The main contributions of this work are summarized as follows: (a) we propose an unsupervised method for extractive multi-document summarization based on the centroid approach and the sentence embedding representations. (b) we adopt three metrics for sentence scoring and show that combining these three metrics leads to significant improvements. (c) we assess the performance of nine sentence embedding models on text summarization task using the standard DUC'2002-2004 datasets. (d) we compare our method with sveral state-of-the-art methods on DUC'2004 and Multi-News datasets.

We performed an extensive experimental analysis to validate the robustness of the proposed multi-document summarization method. In particular, several experiments were conducted on DUC'2002-2004 datasets to evaluate each sentence embedding model exploited in our method. The experimental results showed that the use of sentence embedding representations has considerably improved the results. Additionally, according to ROUGES scores, the results demonstrate that the top-5 performing sentence embedding models for extractive multi-document summarization are: the non-parameterized model uSIF, and the parameterized models USE-DAN, USE-Transformer, NNLM, and InferSent_GloVe. Furthermore, we compared our method with several unsupervised state-of-the-art multi-document summarization methods including three centroid-based methods (Centroid_BOW, Centroid_WordEmbedding_v1, and Centroid_WordEmbedding_v2) and recent supervised deep learning based methods. The obtained results show that our method has significantly outperformed the Centroid_BOW and Centroid_WordEmbedding_v1 methods while it has achieved comparable results to the best performing methods (DPP, ISCISum, ConceptBased_ILP, and Centroid_WordEmbedding_v2). Finally, the obtained results on DUC'2004 and Multi-News datasets show that our method has achieved promising performances compared to the recent deep learning based methods.

The use of sentence embedding representations in extractive multi-document summarization has been shown to be effective. Thus, in future work, we plan to investigate the efficiency of these representations on other summarization tasks such as query-focused text summarization. Additionally, we plan to fine-tune BERT on intermediate tasks from the GLUE benchmark for multi-document summarization. This idea is motivated by the fact that sentence embedding models that are trained or fine-tuned on the SNLI task have shown a good performance (USE-DAN, USE-Transformer, and InferSent). Furthermore, transfer learning using the pre-trained sentence embedding models can be applied to other NLP tasks such as the Word Sense Disambiguation (Correa Jr et al., 2018) as they carry rich contextualized word representations.

Acknowledgments

We would like to thank Professors Rinaldo Lima (Federal University of Pernambuco, Recife, Brazil) and Noureddine En-nahnahi (Laboratory of Informatics and Modeling, FSDM, Sidi Mohammed Ben Abdellah University, Fez, Morocco) for their relevant insights. We also acknowledge the National Institute of Standards and Technology (NIST) for providing us the DUC datasets.

References

Amancio, D.R., Nunes, M.G., Oliveira Jr, O.N., Costa, L.d.F., 2012. Extractive summarization using complex networks and syntactic dependency. Physica A: Statistical Mechanics and its Applications 391, 1855–1864.

Aries, A., Zegour, D.E., Hidouci, W., 2019. Automatic text summarization: What has been done and what has to be done. CoRR abs/1904.00688.

Arora, S., Liang, Y., Ma, T., 2016. A simple but tough-to-beat baseline for sentence embeddings .

Baralis, E., Cagliero, L., Mahoto, N., Fiori, A., 2013. Graphsum: Discovering correlations among multiple terms for graph-based summarization. Information Sciences 249, 96–109.

Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C., 2003. A neural probabilistic language model. Journal of machine learning research 3, 1137–1155.

Bowman, S.R., Angeli, G., Potts, C., Manning, C.D., 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.

Cao, Z., Wei, F., Li, S., Li, W., Zhou, M., Houfeng, W., 2015. Learning summary prior representation for extractive summarization, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pp. 829–833.

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., 2018. Universal sentence encoder for English , 169–174.

Cheng, J., Lapata, M., 2016. Neural summarization by extracting sentences and words , 484–494.

Cho, K., van Merriënboer, B., Bahdanau, D., Bengio, Y., 2014. On the properties of neural machine translation: Encoder–decoder approaches, in: Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pp. 103–111.

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, EMNLP, pp. 670–680.

Conroy, J.M., Goldstein, J., Schlesinger, J.D., O'leary, D.P., 2004. Left-brain/right-brain multi-document summarization, in: In Proceedings of the Document Understanding Conference (DUC'2004).

Correa Jr, E.A., Lopes, A.A., Amancio, D.R., 2018. Word sense disambiguation: A complex network approach. Information Sciences 442, 103–113.

Denil, M., Demiraj, A., De Freitas, N., 2014. Extraction of salient sentences from labelled documents. preprint arXiv:1412.6815 .

Dernoncourt, F., Ghassemi, M., Chang, W., 2018. A repository of corpora for summarization, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC).

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding , 4171–4186.

Edmundson, H.P., 1969. New methods in automatic extracting. Journal of the ACM (JACM) 16, 264–285.

Erkan, G., Radev, D.R., 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research 22, 457–479.

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.

Ferreira, R., de Souza Cabral, L., Lins, R.D., e Silva, G.P., Freitas, F., Cavalcanti, G.D., Lima, R., Simske, S.J., Favaro, L., 2013. Assessing sentence scoring techniques for extractive text summarization. Expert systems with applications 40, 5755–5764.

Garcia, R., Lima, R., Espinasse, B., Oliveira, H., 2018. Towards coherent single-document summarization: an integer linear programming-based approach, in: Proceedings of the 33rd Annual ACM Symposium on Applied Computing, pp. 712–719.

Gehrmann, S., Deng, Y., Rush, A., 2018. Bottom-up abstractive summarization, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4098–4109.

Ghalandari, D.G., 2017. Revisiting the centroid-based method: A strong baseline for multi-document summarization, in: Proceedings of the Workshop on New Frontiers in Summarization, NFiS@EMNLP, pp. 85–90.

Gillick, D., Favre, B., Hakkani-Tür, D., 2008. The ICSI summarization system at TAC 2008, in: Proceedings of the First Text Analysis Conference, TAC.

Hartigan, J.A., Wong, M.A., 1979. Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) 28, 100–108.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

Hokamp, C., Ghalandari, D.G., Pham, N.T., Glover, J., 2020. Dyne: Dynamic ensemble decoding for multi-document summarization. arXiv:2006.08748.

Hong, K., Conroy, J.M., Favre, B., Kulesza, A., Lin, H., Nenkova, A., 2014. A repository of state of the art and competitive baseline summaries for generic news summarization., in: In proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, pp. 1608–1616.

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, M.K., 2017. Extractive text summarization using word vector embedding, in: 2017 International Conference on Machine Learning and Data Science (MLDS), pp. 51–55.

Jin, H., Wang, T., Wan, X., 2020. Multi-granularity interaction network for extractive and abstractive multi-document summarization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6244–6254.

Joshi, A., Fidalgo, E., Alegre, E., Fernández-Robles, L., 2019. Summcoder: An unsupervised framework for extractive text summarization based on deep auto-encoders. Expert Systems with Applications 129, 200–215.

Kågebäck, M., Mogren, O., Tahmasebi, N., Dubhashi, D., 2014. Extractive summarization using continuous vector space models, in: Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC), pp. 31–39.

Kalchbrenner, N., Grefenstette, E., Blunsom, P., 2014. A convolutional neural network for modelling sentences, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 655–665.

Kim, Y., 2014. Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751.

Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S., 2015. Skip-thought vectors, in: Advances in neural information processing systems, pp. 3294–3302.

Kobayashi, H., Noguchi, M., Yatsuka, T., 2015a. Summarization based on embedding distributions, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP, pp. 1984–1989.

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

Kulesza, A., Taskar, B., et al., 2012. Determinantal point processes for machine learning. Foundations and Trends® in Machine Learning 5, 123–286.

Le, Q.V., Mikolov, T., 2014. Distributed representations of sentences and documents, in: Proceedings of the 31th International Conference on Machine Learning, ICML, pp. 1188–1196.

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.

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

Lin, H., Bilmes, J., 2011. A class of submodular functions for document summarization, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 510–520.

McDonald, R., 2007. A study of global inference algorithms in multi-document summarization, in: European Conference on Information Retrieval, pp. 557–564.

Metzler, D., Kanungo, T., 2008. Machine learned sentence selection strategies for query-biased summarization, in: In Sigir learning to rank workshop, pp. 40–47.

Mihalcea, R., Tarau, P., 2004. Textrank: Bringing order into text, in: Proceedings of the 2004 conference on empirical methods in natural language processing, pp. 404–411.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, pp. 3111–3119.

Mohd, M., Jan, R., Shah, M., 2020. Text document summarization using word embedding. Expert Systems with Applications 143, 112958.

Oliveira, H., Ferreira, R., Lima, R., Lins, R.D., Freitas, F., Riss, M., Simske, S.J., 2016. Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization. Expert Systems with Applications 65, 68–86.

de Oliveira, H.T.A., Lins, R.D., Lima, R., de Freitas, F.L.G., Simske, S.J., 2018. A concept-based ilp approach for multi-document summarization exploring centrality and position. 2018 7th Brazilian Conference on Intelligent Systems (BRACIS) , 37–42.

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.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2227–2237.

Radev, D.R., Jing, H., Styś, M., Tam, D., 2004. Centroid-based summarization of multiple documents. Information Processing & Management 40, 919–938.

Ramos, J., et al., 2003. Using tf-idf to determine word relevance in document queries, in: In Proceedings of the first instructional conference on machine learning, pp. 133–142.

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.

Saggion, H., Poibeau, T., 2013. Automatic text summarization: Past, present and future, in: Multi-source, multilingual information extraction and summarization, pp. 3–21.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C., 2013. Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the 2013 conference on empirical methods in natural language processing, pp. 1631–1642.

Steinberger, J., Jezek, K., 2004. Using latent semantic analysis in text summarization and summary evaluation. Proceedings of the 7th International Conference ISIM'04 , 93–100.

Tohalino, J.V., Amancio, D.R., 2018. Extractive multi-document summarization using multilayer networks. Physica A: Statistical Mechanics and its Applications 503, 526–539.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in neural information processing systems, pp. 5998–6008.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S., 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding, in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pp. 353–355.

Yang, Z., Zhu, C., Chen, W., 2018. Zero-training sentence embedding via orthogonal basis. preprint arXiv:1810.00438 .

Yao, J.g., Wan, X., Xiao, J., 2017. Recent advances in document summarization. Knowledge and Information Systems 53, 297–336.

Yasunaga, M., Zhang, R., Meelu, K., Pareek, A., Srinivasan, K., Radev, D., 2017. Graph-based neural multi-document summarization. preprint arXiv:1706.06681 .

Yin, W., Pei, Y., 2015. Optimizing sentence modeling and selection for document summarization, in: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI, pp. 1383–1389.

Yogatama, D., Liu, F., Smith, N.A., 2015. Extractive summarization by maximizing semantic volume, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1961–1966.

Yousefi-Azar, M., Hamey, L., 2017. Text summarization using unsupervised deep learning. Expert Systems with Applications 68, 93–105.

Zhang, C., Sah, S., Nguyen, T., Peri, D., Loui, A., Salvaggio, C., Ptucha, R., 2017. Semantic sentence embeddings for paraphrasing and text summarization, in: 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 705–709.

Zhao, H., Lu, Z., Poupart, P., 2015. Self-adaptive hierarchical sentence model, in: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI, pp. 4069–4076.

Zhong, S.h., Liu, Y., Li, B., Long, J., 2015. Query-oriented unsupervised multi-document summarization via deep learning model. Expert systems with applications 42, 8146–8155.

A. Example Summaries

Table 8 illustrate example of generated summaries by Run 1, Run 2, Run 3, and Run 4 for DUC2004 using uSIF sentence embedding model as well as the reference summaries of the same cluster of documents.

Table 8
Example of generated summaries by Run 1, Run 2, Run 3, Run 4 along with the reference summaries.

Run 1

• The Yugoslav war crimes tribunal Monday acquitted a Muslim military commander of war crimes against Bosnian Serb prisoners in 1992, but convicted three underlings in the first U.N. case dealing with anti-Serb atrocities.

• In its first case to deal with atrocities against Serbs during Bosnia's civil war, a U.N. war crimes tribunal on Monday convicted three prison officials and guards, but acquitted a top military commander who oversaw the facility.

• The Yugoslav war crimes tribunal cleared Zejnil Delalic, a Muslim, of responsibility for war crimes committed against Serb captives at a Bosnian government-run prison camp under his command.

• Set up in 1993, the U.N. court has convicted three Muslims, two Bosnian Croats and a Bosnian Serb of war crimes including murder, rape and torture, but it has yet to register a genocide conviction.

Run 2

• In its first case to deal with atrocities against Serbs during Bosnia's civil war, a U.N. war crimes tribunal on Monday convicted three prison officials and guards, but acquitted a top military commander who oversaw the facility.

• The Yugoslav war crimes tribunal Monday acquitted a Muslim military commander of war crimes against Bosnian Serb prisoners in 1992, but convicted three underlings in the first U.N. case dealing with anti-Serb atrocities.

• American and allied forces in Bosnia on Wednesday arrested a Bosnian Serb general who was charged with genocide by the international war crimes tribunal in a recent secret indictment.

• Some of the closest combat in the half year of the Kosovo conflict, to the point of fighting room to room and floor to floor, occurred near this village six weeks ago, in the days before 21 women, children and elderly members of the Delijaj clan were massacred by Serbian forces, their mutilated bodies left strewn on the forest floor.

Run 3

• The Yugoslav war crimes tribunal Monday acquitted a Muslim military commander of war crimes against Bosnian Serb prisoners in 1992, but convicted three underlings in the first U.N. case dealing with anti-Serb atrocities.

• In its first case to deal with atrocities against Serbs during Bosnia's civil war, a U.N. war crimes tribunal on Monday convicted three prison officials and guards, but acquitted a top military commander who oversaw the facility.

• Set up in 1993, the U.N. court has convicted three Muslims, two Bosnian Croats and a Bosnian Serb of war crimes including murder, rape and torture, but it has yet to register a genocide conviction.

Run 4

• The Yugoslav war crimes tribunal Monday acquitted a Muslim military commander of war crimes against Bosnian Serb prisoners in 1992, but convicted three underlings in the first U.N. case dealing with anti-Serb atrocities.

• Yugoslavia must cooperate with the U.N. war crimes tribunal investigating alleged atrocities during the wars in Croatia and Bosnia, international legal experts meeting in Belgrade said Sunday.

• He is accused of directing the attack on Srebrenica in 1995, one of the most chilling and influential events of the Bosnian war, when some 7,000 Bosnian Muslim men were marched off, presumably to their deaths, as U.N. peacekeepers stood by.

Reference summary 1

• The UN war crimes tribunal demands Yugoslavia's full cooperation in its investigations. • It blasted Belgrade for refusing to let investigators probe alleged atrocities in Kosovo.

• The tribunal acquitted a Muslim commander, but convicted 3 underlings.

• The commander was greeted by hundreds at Sarajevo a irport on his return.

• Survivors say Serb forces in Kosovo took revenge against civilians for their battle losses.

• Trial begins for Bosnian Serb Jelisic, nicknamed "Serb Adolf," who boasted of many killings.

• He is accused of genocide, although he confessed to 12 killings.

• U.S. forces in Bosnia arrested a Serb general accused of genocide at Srebrenica.

Reference summary 2

• Yugoslavia has cooperated with the UN War crimes tribunal in cases where Serbs were victims in Bosnia and Croatia, but has been slow to allow investigation of alleged atrocities in Kosovo.

• Serbs fighting there suffered losses to the guerillas and took revenge on civilians including women and children.

• The war crimes tribunal acquitted a Muslim military commander of war crimes against Bosnian Serb prisoners but convicted three underlings.

• A Bosnian Serb, known as "Serb Adolf", accused of genocide, admitted killing Muslims and Croats.

• Allied forces arrested a Bosnian Serb general charged with genocide and who could implicate Slobodan Milosevic.

Reference summary 3

• Yugoslavia was told to cooperate with the UN War Crimes Tribunal whether Serbs were victims or accused.

• Belgrade refused visas to Kosovo atrocity investigators.

• Serbians took revenge on Kosovo civilians for heavy losses.

• Muslim commander Delalic was acquitted of anti-Bosnian Serb atrocities after 3 years and welcomed home. • 3 of his underlings were convicted.

• Croat commander Mucic was convicted for permitting atrocities.

• Serb Adolf Goran Jelisic was freed from jail by Bosnian Serbs and told to go kill Muslims. • He confessed to murdering 12.

• Maj. Gen. Krstic, a Bosnian Serb, is the first serving officer to be arrested. • He directed the attack on Srebrenica.

Reference summary 4

• Milosevic cooperates with the U.N. war crimes tribunal when Serbs are victims, but is an obstructionist when they are the accused.

• Officials, for example, limited U.N. investigators' access to Kosovo, where Serbs massacred 21 ethnic Albanian civilians of the Delijaj clan.

• Further, while U.S. and allied forces arrested Bosnian Serb General Krstic on genocide charges, other indicted, high-ranking Serb leaders are protected in Serbia.

• Meanwhile, in the first trial involving anti-Serb acts, a Muslim military commander was freed, but three prison camp officials were convicted.

• Also, in The Hague, the genocide trial of Goran Jelisic, the "Serb Adolf," has begun.

HIGHLIGHTS

- An unsupervised method for extractive multi-document summarization
- Pre-trained sentence embedding models are used for sentences representations
- Centroid approach is applied to compute the sentence content relevance score
- Sentence selection based on sentence relevance, novelty and position scores
- The use of sentence embedding methods leads to significant improvements

Credit Author Statement

1. **Salima Lamsiyah**: Conceptualization, Methodology, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Visualization, Investigation, Software.
2. **Abdelkader El Mahdaouy**: Conceptualization, Methodology, Formal analysis, Writing - Review & Editing, Software.
3. **Bernard Espinasse** : supervision, Validation, Writing - Original Draft, Resources.
4. **Saïd Ouatik El Alaoui** : supervision, Validation, Resources.

Conflicts of Interest Statement

Manuscript title: *An Unsupervised Method for Extractive Multi-Document Summarization based on Centroid Approach and Sentence Embeddings*

The authors whose names are listed immediately below certify that this manuscript has not been submitted to, nor is under review at, another journal or other publishing venue. All the authors have participated in improving this manuscript. And, the authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

Details of the authors:

1. Salima Lamsiyah * **(corresponding author)**
   Laboratory of Engineering Sciences, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco.
   Laboratory of Informatics, Signals, Automatic, and Cognitivism, FSDM, Sidi Mohamed Ben Abdellah University, Fez, Morocco.
   Email: salimalamsiyah@gmail.com

2. Abdelkader El Mahdaouy
   Laboratory of Informatics, Signals, Automatic, and Cognitivism, FSDM, Sidi Mohamed Ben Abdellah University, Fez, Morocco.
   Email: abdelkader.elmahdaouy@usmba.ac.ma

3. Bernard Espinasse
   Aix-Marseille Université, Université de Toulon, LIS UMR CNRS 7020,
   Marseille, France
   Email: bernard.espinasse@lis-lab.fr

4. Saïd El Alaoui Ouatik
   Laboratory of Engineering Sciences, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco.
   Laboratory of Informatics, Signals, Automatic, and Cognitivism, FSDM, Sidi Mohamed Ben Abdellah University, Fez, Morocco.
   Email: said.ouatikelalaoui@usmba.ac.ma