A Spectral Method for Unsupervised Multi-Document Summarization

Kexiang Wang[1], **Baobao Chang**[1,2] and **Zhifang Sui**[1,2]
[1]Key Laboratory of Computational Linguistics, Ministry of Education,
School of Electronics Engineering and Computer Science, Peking University, Beijing, China
[2]Peng Cheng Laboratory, Guangdong, China
{wkx,chbb,szf}@pku.edu.cn

Abstract

Multi-document summarization (MDS) aims at producing a good-quality summary for several related documents. In this paper, we propose a spectral-based hypothesis, which states that the goodness of summary candidate is closely linked to its so-called spectral impact. Here spectral impact considers the perturbation to the dominant eigenvalue of affinity matrix when dropping the summary candidate from the document cluster. The hypothesis is validated by three theoretical perspectives: semantic scaling, propagation dynamics and matrix perturbation. According to the hypothesis, we formulate the MDS task as the combinatorial optimization of spectral impact and propose an accelerated greedy solution based on a surrogate of spectral impact. The evaluation results on various datasets demonstrate: (1) The performance of the summary candidate is positively correlated with its spectral impact, which accords with our hypothesis; (2) Our spectral-based method has a competitive result as compared to state-of-the-art MDS systems.

1 Introduction

Given a cluster of documents related to the same topic or event, the task of multi-document summarization (MDS) centers on a brief summary of the cluster. As emphasized by Lebanoff et al. (2018), for this task, the labeled training data (i.e. cluster-summary pairs) are scarce. Hence dealing with it in an unsupervised paradigm becomes a reasonable choice. For the unsupervised MDS task, the automatic summarizer is required to discover the main content of the document cluster without the guidance of golden summaries. To preserve the fluency and grammaticality of summary, we mainly focus on the extractive method in which summary sentences are extracted from the original document cluster.

In this paper, we propose a novel spectral-based hypothesis for the unsupervised MDS task. The hypothesis states that the goodness (or effectiveness) of any summary candidate is closely linked with its spectral impact on the document cluster. The spectral impact of a summary candidate quantifies the perturbation to the dominant eigenvalue (in modulus) of affinity matrix when dropping the candidate from the document cluster. In other words, the hypothesis points out the spectral impact as an indicator of the MDS task, which is the first attempt to characterize MDS from a spectral viewpoint explicitly. As a representation of the document cluster, the affinity matrix supports the definition of spectral impact. Adjusting the building of the affinity matrix can bring out the best in the hypothesis. To validate the proposed hypothesis, we provide both theoretical explanations and empirical evidence. *Theoretically*, the spectral impact caused by dropping a summary from the cluster can be characterized from three different perspectives (see §2.4). *Empirically*, for any summary candidate, the real dataset witnesses a positive correlation between its performance and computed spectral impact. For a particular MDS task, applying the hypothesis leads to a constrained optimization problem where the objective function is spectral impact. Our summarizer utilizes an accelerated greedy algorithm based on a surrogate of spectral impact. The competitive results of our summarizer have been obtained on various datasets.

The differences between prior works and our method are clarified for unsupervised MDS:

(1) *Underlying hypothesis*. The hypothesis indicates the mechanism for the summarization. For instance, manifold-ranking-based methods share the hypothesis that a good summary sentence has a high ranking on the low-dimensional manifold that documents reside in (Wan et al., 2007; Cheng et al., 2011; Li et al., 2011). However, the reasonableness

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 435–445,
November 16–20, 2020.

of this manifold hypothesis has not been directly evaluated. Another hypothesis in the sparse-coding-based methods (Li et al., 2015b; Liu et al., 2015; Yao et al., 2015) regards the original sentences as a linear combination of summary sentences. This leads to an intuitive reconstruction, whereas linear combination is more a simplification than a necessity. Our proposed hypothesis offers a spectral viewpoint and will be explicitly validated on the real dataset.

(2) *Optimization objective*. Multi-criteria optimization is suitable for MDS as various criteria (goals) exist in the task, such as relevancy criterion and non-redundancy criterion. For instance, Lin and Bilmes (2011) is a bi-criteria case that imposes the submodularity constraint on each criterion. Multi-criteria loss functions in neural-network-based methods (Ma et al., 2016; Chu and Liu, 2019; Zheng et al., 2019) include the reconstruction errors from different spaces. In the above cases, the overall objective functions used include some hyperparameters for gluing singletons. Comparatively, our proposed objective (spectral impact) has a compact form. It avoids the hyperparameter setting and simulates the non-separable processing of multiple MDS criteria by human beings.

(3) *Model complexity*. There is a trade-off between model complexity and model interpretability. For instance, the reported performance of the aforementioned deep-neural-based models is elusive, and there exists no general principle to further improve them. Our summarizer realizes the interpretable behavior based on verified hypothesis while preserving enough model complexity by the flexible affinity matrix (as a plug-in).

Our main contributions are twofold: (1) A novel spectral-based hypothesis for unsupervised MDS, which gains support from both theoretical and empirical sides; (2) An accelerated greedy algorithm for solving the hypothesis-driven optimization problem.

The rest of the paper is organized as follows. §2 gives the details of our method, including the spectral-based MDS hypothesis and the greedy algorithm to solve the spectral optimization problem. Evaluation results, related work and conclusions are covered in §3, §4 and §5 respectively.

2 Spectral-based MDS

What role does the summary play in the process of MDS? Our proposed hypothesis offers a spectral insight and brings out a workable formulation of MDS.

2.1 Notations

We use calligraphic fonts for sets, capital bold letters for matrices and lower-case bold letters for vectors. The universal set $\mathcal{C}$ is formed by splitting and gathering the sentences from document cluster, i.e. $\mathcal{C}=\{s_1, s_2, ..., s_n\}$ (s_i represents the i-th sentence and n is the total number of sentences). Each sentence has its ordinal number, e.g. o_i of sentence s_i indicates it is the (o_i)-th sentence in the document that s_i belongs to. The summary candidate (subset of $\mathcal{C}$) is denoted as $\mathcal{S}$. We represent the affinity matrix of document cluster as: $\mathbf{A}=\{a_{ij}\}_{n\times n}$. In addition, the dominant eigenvalue (in modulus) and the corresponding eigenvector of $\mathbf{A}$ are denoted as $\lambda(\mathbf{A})$ and $\mathbf{v}$, respectively.

Dropping a set from a matrix: emptying all the rows and columns whose indexes occur in the set. Consider the operation of dropping $\mathcal{S}$ from $\mathbf{A}$. If we denote the operation itself and the resultant matrix as $\mathbf{A}\backslash\mathcal{S}$, then

$$\mathbf{A}\backslash\mathcal{S} = \begin{cases} 0, & s_i \in \mathcal{S} \text{ or } s_j \in \mathcal{S}, \\ a_{ij}, & \text{otherwise.} \end{cases}$$

2.2 Spectral Hypothesis

When representing the document cluster as a matrix, the matrix spectrum (i.e. a collection of eigenvalues) can uncover its different facets. Note that the dominant eigenvalue especially corresponds to the key facet, which gives a clue as to the main content that the summarizer needs to discover. For the extractive MDS, we propose the spectral-based hypothesis:

> GIVEN: Affinity matrix $\mathbf{A}$, the matrix representation of document cluster; set $\mathcal{S}$, any summary candidate including some original sentences.
> DEFINITION: Spectral impact of $\mathcal{S}$ is the perturbation to dominant eigenvalue of $\mathbf{A}$ when dropping $\mathcal{S}$ from $\mathbf{A}$, i.e. $\Delta\lambda(\mathcal{S}) \triangleq \lambda(\mathbf{A}) - \lambda(\mathbf{A}\backslash\mathcal{S})$.
> HYPOTHESIS: Goodness of $\mathcal{S}$ as a summary has a close link with its spectral impact $\Delta\lambda(\mathcal{S})$.

The above hypothesis tells us that the goodness (effectiveness) of any summary candidate can be determined by the proposed spectral impact, which reflects the change of dominant eigenvalue when the summary candidate $\mathcal{S}$ is left out. Affinity matrix $\mathbf{A}$ supports the definition of spectral impact, which

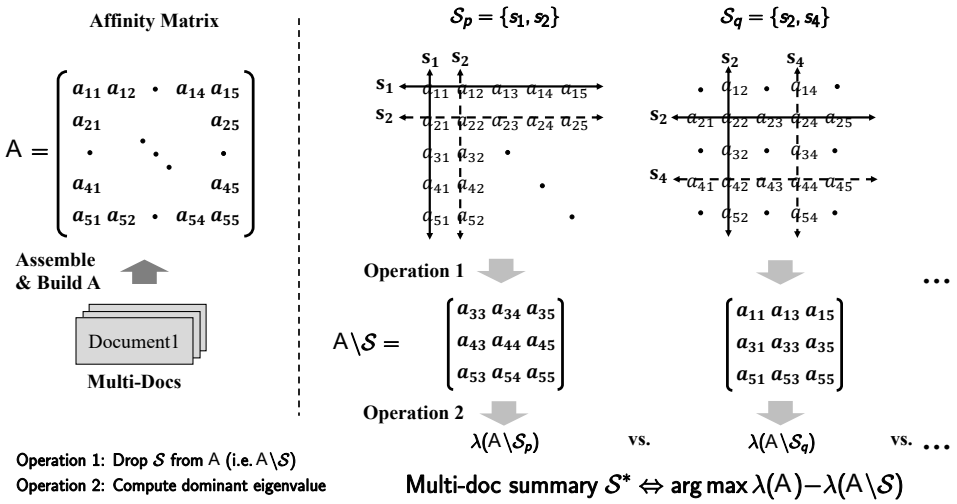

Figure 2.1: An example depicting the spectral-based hypothesis for the task of unsupervised MDS ($n=5$, $k=2$). The hypothesis suggests using spectral impact to judge whether a summary candidate is good or not.

stores the pairwise affinity of sentences as the name suggests.

For a specific MDS task, applying the hypothesis leads to an optimization formulation as follows:

$$\mathcal{S}^* = \arg\max_{\mathcal{S}\subseteq\mathcal{C}} \Delta\lambda(\mathcal{S}), \quad \text{s.t. } |\mathcal{S}| \leqslant k. \quad (1)$$

The set $\mathcal{C}$ denotes the universal set and the number k specifies the maximum capacity of candidate $\mathcal{S}$.

The above formulation sets the spectral impact to be the objective function. The inherent rationality can be verified partially by these properties:

(a) *monotonicity*: $\Delta\lambda(\mathcal{S}_1) \leqslant \Delta\lambda(\mathcal{S}_2)$ for any $\mathcal{S}_1 \subseteq \mathcal{S}_2$ (see Li et al., 2012, Theorem 1);

(b) *normalization*: $\Delta\lambda(\Phi) = 0$, $\Delta\lambda(\mathcal{C}) = \lambda(\mathbf{A})$ (Φ denotes empty set).

In the context of MDS, property (a) points out that a whole summary has more goodness than its components and property (b) regulates a reasonable range of the goodness of any summary candidate.

An overview of our hypothesis can be found in Figure 2.1. The document cluster with its matrix representation **A** is depicted in the left part, while the right part gives the dropping operation and the computation of spectral impact for two summary candidates $\mathcal{S}_p$ and $\mathcal{S}_q$. The goodness of each candidate is judged by their spectral impacts, and the winner $\mathcal{S}^*$ stands out with the largest spectral impact.

Notice that cardinality constraint is adopted in Problem (1) to specify the length limit of summary. Other reasonable constraints are also available, such as the knapsack constraint and the non-uniform matroid constraint (Welsh, 1976). The relevant conclusions and solutions discussed in the following sections continue to be applicable for those constraints.

2.3 Affinity Matrix

Many prior works have adopted **A** for the MDS task, such as Yang et al. (2018) and Yang et al. (2019). Each element in the affinity matrix **A** is a pairwise affinity of two different sentences. Since our hypothesis depends on **A**, a better MDS performance can be expected by adjusting the building of **A**. Sentence embeddings play a vital role in the process of building **A**, since affinity a_{ij} can be set to be the cosine similarity of the embeddings of sentences s_i and s_j (i.e. $a_{ij} = a_{ji}$ and $a_{ii} = 0$). For comparison purposes, we consider the following three strategies of building sentence embeddings.

Tf-isf: the simple tf-idf model with a finer granularity. More details can be found in Wan et al. (2007) and Wang et al. (2017).

ESE: the enhanced feature embedding model (Yang et al., 2019). The embedding of each sentence is the concatenation of all components: paragraph vector, positional embedding and three feature embeddings (namely word-part-of-speech, bigram and trigram).

BERT: the sentence encoder that learns vector representations by pre-training a deep bi-directional Transformer network (Devlin et al., 2019). The advantage is that BERT is context-sensitive when considering the word embedding.

Notice that the leading sentences in each document should have priority in the summary extraction. For injecting this knowledge, a_{ij} is multiplied

by the average positional weight $1/(o_i+o_j)$. This can differentiate the sentences across documents and preserve the symmetry of $\mathbf{A}$.

2.4 Justifications of Hypothesis

We validate our spectral-based hypothesis by the following three complementary perspectives:

Semantic scaling: *dominant eigenvalue of affinity matrix determines the vector scaling in semantic space*. The n-dimensional semantic space is constructed as follows: Each sentence in the document cluster represents a different dimension and the i-th axis of the space corresponds to the i-th sentence. Then the affinity matrix $\mathbf{A}_{n\times n}$ can be seen as a linear operator on this semantic space and the pairwise affinity a_{ij} regulates the transformation between the i-th axis and j-th axis. Given an arbitrary nonzero vector $\mathbf{x}$ in the space, the transformed vector is $\mathbf{Ax}$. Then the property holds:

$$\|\mathbf{Ax}\| \leqslant \lambda(\mathbf{A})\|\mathbf{x}\|. \tag{2}$$

Notation $\|\cdot\|$ denotes the Euclidean norm of vector. This is a sharp bound as equality holds only if $\mathbf{x}$ is the dominant eigenvector of operator $\mathbf{A}$.

The property shows that the scaling up of any vector (namely $\|\mathbf{Ax}\|/\|\mathbf{x}\|$) is not larger than $\lambda(\mathbf{A})$. Hence, the dominant eigenvalue $\lambda(\mathbf{A})$ characterizes the ability of operator $\mathbf{A}$ to scale up any vector in the semantic space that document cluster resides in. When dropping the summary candidate $\mathcal{S}$, the transformations to and from all axes covered by $\mathcal{S}$ will no longer exist for operator $\mathbf{A}$. In other words, there is no contribution for scaling up vectors from the i-th axis for any sentence $s_i \in \mathcal{S}$. When the best-quality summary is dropped, the main components of operator $\mathbf{A}$ are emptied, which causes the largest reduction of its ability to scale up vectors. Therefore, the dominant eigenvalue, indicator of this ability, can be used to locate the multi-document summary, as proposed in our hypothesis.

Propagation dynamics: *isolating the summary blocks the information dynamics*. In this perspective, there is a spread of information over the document cluster according to the underlying network specified by matrix $\mathbf{A}$. The pairwise affinity a_{ij} indicates the propagation rate between sentences s_i and s_j (more similar they are, more rapid the propagation occurs). The question that arises here is whether the information propagated from a few seed sentences will form a pandemic or become extinct in the long term. In epidemiology, the virus (information) will form a pandemic only if the basic reproduction number R_0 of this virus is larger than 1 (Jones, 2007). For instance, R_0 of COVID-19 is about 3.28 (> 1) (Liu et al., 2020), which uncovers the inevitable propagation of this virus.

Many works (Wang et al., 2003; Prakash et al., 2012; Chen, 2018) have found out that R_0 is proportional to the dominant eigenvalue of the underlying information network. Thus a small dominant eigenvalue corresponds to a small value of R_0, which hinders the information propagation. For the MDS task, when isolating the best-quality summary (i.e. $\mathbf{A}\backslash\mathcal{S}$), the remainder of the document cluster will become the hardest for information propagation. Our hypothesis is consistent with this finding as the summary $\mathcal{S}$ found by solving Problem (1) is able to reduce $\lambda(\mathbf{A}\backslash\mathcal{S})$ the most.

Matrix perturbation: *spectral impact considers both the relevancy and non-redundancy goal of MDS*. For analyzing the behavior of spectral impact, we expand it using first-order matrix perturbation theory (Stewart, 1990) as follows:

$$\begin{aligned}\Delta\lambda(\mathcal{S}) &= \mathbf{u}^{'}\mathbf{Eu} + \mathcal{O}(\|\mathbf{E}\|^2) \\ &= 2\sum_{s_i\in\mathcal{S}} {u_i}^2\lambda(\mathbf{A}) - \sum_{s_i,s_j\in\mathcal{S}} u_i a_{ij} u_j + \mathcal{O}(\|\mathbf{E}\|^2),\end{aligned} \tag{3}$$

where $\mathbf{E} = \mathbf{A} - \mathbf{A}\backslash\mathcal{S}$, $\mathbf{Au} = \lambda(\mathbf{A})\mathbf{u}$, $\|\mathbf{u}\| = 1$.

Let us analyze each term of the expansion shown in Eq. (3). The first sums up the score of $2{u_i}^2\lambda(\mathbf{A})$ for any sentence $s_i \in \mathcal{S}$. The value of u_i is a measure for the relevancy of sentence s_i, since eigenvector centrality has been typically used for ranking sentences (Erkan and Radev, 2004; Bellaachia and Al-Dhelaan, 2014; Al-Dhelaan, 2015). Hence, the first term is an indicator for the **relevancy** of summary $\mathcal{S}$. The second term is a penalty that considers every pair of summary sentences. Specifically, the penalty is $u_i a_{ij} u_j$ for sentences s_i and s_j. When a_{ij} is large (sentences are redundant), the penalty becomes prominent. Thus the second term measures the **non-redundancy** of summary $\mathcal{S}$. The third term $\mathcal{O}(\|\mathbf{E}\|^2)$ is relatively small compared to the preceding two because matrix $\mathbf{E}$ is nearly dominated by zeros ($|\mathcal{S}| \ll n$). Hence the third term will not change the main behavior of $\Delta\lambda(\mathcal{S})$.

2.5 Algorithm

The naive idea for solving Problem (1) is to enumerate all possible combinations and find the best summary. The time complexity is $\binom{n}{k}n^2$ if it takes $\mathcal{O}(n^2)$ time to compute the dominant eigenvalue

of matrix (say using the method proposed by Lanczos (1950)). This exact enumeration algorithm is *infeasible* even when n is 500 and k is 5.

Theoretically, Problem (1) falls into spectral optimization that has been proved to be NP-hard in many cases (Van Mieghem et al., 2011). To avoid the time-consuming eigen-decomposition, some works resort to the QR decomposition of matrix (Li et al., 2015a; Chen et al., 2018). However, to actually compute QR decomposition, they depend on the Gram–Schmidt process that is inherently numerically unstable, which impedes the optimization process. In this paper, we bypass all these matrix decomposition and propose a straightforward surrogate for spectral impact, which is both effective and efficient. Based on a bound for dominant eigenvalue of $\mathbf{A}\backslash\mathcal{S}$ (Theorem 2.14 in Stevanovic (2014)), the surrogate is proposed as follows:

Surrogate of spectral impact:

$$\Delta\lambda(\mathcal{S}) \approx \frac{\sum_{s_i\in\mathcal{S}} v_i^2\lambda - \sum_{s_i,s_j\in\mathcal{S}} v_i a_{ij} v_j}{\sum_{s_i\in\mathcal{C}} v_i^2 - \sum_{s_i\in\mathcal{S}} v_i^2}$$

where λ is the dominant eigenvalue $\lambda(\mathbf{A})$ and $\mathbf{v}$[1] is its corresponding eigenvector of matrix $\mathbf{A}$.

By using the surrogate for acceleration, we consider a greedy strategy to iteratively select $\mathcal{S}$, which is listed in Alg. 2.1. First, we compute the dominant eigenvalue λ and eigenvector $\mathbf{v}$ of $\mathbf{A}$ (line 1). At each iteration (lines 3 to 7), the sentence s_τ maximizing the marginal gain of $\Delta\lambda(\mathcal{S})$ is extracted based on the previously selected set $\mathcal{S}$ (i.e. maximizing $\Delta\lambda(\mathcal{S}\cup\{s_j\})$–$\Delta\lambda(\mathcal{S})$, line 4) and added to $\mathcal{S}$ (line 5). Also, the auxiliary vector $\mathbf{w}$ and scalar x should be updated according to the numerator and denominator of the surrogate (lines 6, 7). The operator '$\odot$' and '$\cdot$', for any two vectors, are their Hadamard product and inner product, respectively.

The lemma below demonstrates that Alg. 2.1 has a quadratic time complexity, which is evidently better than the exponential one of naive enumeration.

Lemma 2.1. *The time complexity of Alg. 2.1 is* $\mathcal{O}(n^2 + kn)$.

Proof. Computing the dominant eigen-pair of matrix $\mathbf{A}$ takes $\mathcal{O}(n^2)$ time. The initializations of the vector $\mathbf{w}$ and scalar x are both linear time operations, i.e. $\mathcal{O}(n)$. At each iteration, all n sentences need to be traversed for extracting s_τ and updating $\mathbf{w}$. Thus the total complexity is $\mathcal{O}(n^2 + kn)$. □

Algorithm 2.1: Accelerated Spectral MDS

Input: the affinity matrix $\mathbf{A}$ and the budget k
Output: the summary $\mathcal{S}$

1 Compute the dominant eigen-pair $(\lambda, \mathbf{v})$ of $\mathbf{A}$;
2 Initialize: $\mathcal{S} \leftarrow \Phi$, $\mathbf{w} \leftarrow \lambda\mathbf{v} \odot \mathbf{v}$, $x \leftarrow \mathbf{v} \cdot \mathbf{v}$;
3 **for** $i \leftarrow 1 : k$ **do**
4 Let $\tau \leftarrow \arg\max_j \{\frac{w_j}{x - v_j^2} | j \in [1, n]; s_j \notin \mathcal{S}\}$;
5 Add s_τ to $\mathcal{S}$;
6 Update: $w_j \leftarrow w_j - 2v_j a_{j\tau} v_\tau$ for all $s_j \notin \mathcal{S}$;
7 Update: $x \leftarrow x - v_\tau^2$;
8 **return** $\mathcal{S}$

To get the final summary, we reorder the summary sentences returned by Alg. 2.1 according to their positions in the corresponding document.

3 Experiments

3.1 Datasets

Three datasets are selected in the following experiments to provide a complete evaluation of our method. Two domains have been taken into account: news (DUC and Multi-News) and business reviews (Yelp). Table 3.1 lists some key characteristics of these datasets.

DUC 2004[2] (task 2): the DUC task that contains a benchmark dataset. There are 50 document clusters, each of which includes 10 documents about the same news event. In addition, four human-written summaries are offered for each cluster to be the reference (golden) summary.

Yelp[3]: an all-purpose dataset that can be utilized for MDS. We only use the subset that has the reference summary (the test split offered by Chu and Liu (2019)): 100 businesses (document clusters), each of which includes 8 reviews (documents). One reference summary was collected for each cluster using crowdsourcing. More details of building the dataset can be found in Chu and Liu (2019).

Multi-News[4]: a large-scale dataset collected from news aggregator (Fabbri et al., 2019). It has 5622 document clusters (in the test split offered by the original paper), and multiple documents are present

[1] The eigenvector $\mathbf{v}$ can be of arbitrary length, which differs from the normalized vector $\mathbf{u}$ in Eq. (3).

[2] https://duc.nist.gov/duc2004/tasks.html
[3] https://www.yelp.com/dataset
[4] https://github.com/Alex-Fabbri/Multi-News

in each cluster. Furthermore, each cluster is attached with one human-written reference summary.

	DUC 2004	Yelp	Multi-News
Domain	News	Business review	News
#Clusters	50	100	5622
#Docs per cluster	10	8	2∼10
#Ref. per cluster	4	1	1
#Doc sources	2	1	>1500

Table 3.1: Dataset statistics (only showing test split).

3.2 Experimental Details

For the extractive MDS method (including ours), the pre-processing includes paragraph splitting, sentence splitting and word tokenization. In our method, all the splitted sentences are gathered in set $\mathcal{C}$. The input of Alg. 2.1 includes the affinity matrix $\mathbf{A}$ which is built according to the strategies stated in §2.3. Specifically, the strategy utilizing tf-isf vectors has a word bag that contains all the stemmed words found in the dataset (word stemming using Porter's stemmer[5]). For the strategy ESE, we pre-trained all different sentence embeddings on Daily Mail dataset (Hermann et al., 2015) by following the guideline of Yang et al. (2019) (the dimension of concatenated embedding is 800). For the strategy BERT, we used the uncased BERT-Base model[6] pre-trained on Wikipedia, through bert-as-service[7] to obtain the sentence embedding of 768 dimensions. All the experiments are performed on a machine with two CPUs (3.5GHz) and one GPU (16G memory).

The extractive MDS methods need a length limit of summary to terminate the extraction of summary sentences. We adopt 100 words as the length limit in the DUC dataset, instead of 665 bytes specified by the official task. The change has also been made to provide the same setting for evaluating various methods in Hong et al. (2014) and Zheng et al. (2019). For the Yelp dataset, we set the limit to be the 99.5^{th} percentile less than the maximum length of any document; for Multi-News, the limit is set as 300 words. The same settings have been adopted in Chu and Liu (2019) and Fabbri et al. (2019), respectively.

[5]https://tartarus.org/martin/PorterStemmer/
[6]https://github.com/google-research/bert
[7]https://github.com/hanxiao/bert-as-service

3.3 Evaluation Metrics

We adopt ROUGE (Lin, 2004) as the automatic metric, which has been observed in a good agreement with human judgment (Owczarzak et al., 2012). It measures the overlap of N-grams (R-N) and skip-bigrams with a maximum distance of four words (R-SU4). Also, it can be computed based on the longest common subsequence (R-L). Each version of ROUGE has their scores oriented to recall, precision and F1.

In the experiments, we report the different combinations of ROUGE scores for each dataset, which have been recommended and adopted by previous works. Specifically, the recall scores of R-1,2,4 will be reported for the DUC 2004 dataset according to Hong et al. (2014), Wang et al. (2017) and Zheng et al. (2019); the F1 scores of R-1,2,L will be reported for Yelp as in Chu and Liu (2019); the F1 scores of R-1,2,SU4 will be reported for Multi-News as in Fabbri et al. (2019). The toolkit for computing ROUGE metrics is ROUGE-1.5.5[8] and its option is set to be '-m -c 95 -r 1000 -f A -p 0.5 -t 0'.

3.4 Comparing Methods

We compare our method with both traditional and state-of-the-art MDS methods.

Lead: The documents in a cluster are randomly shuffled, and the first sentence of the document is added to the summary until the length limit is reached.

LexRank (Erkan and Radev, 2004): It performs the sentence relevancy estimation by the random walk process on the sentence graph.

CLASSY04 (Conroy et al., 2004): It ranked first in the official evaluation of DUC 2004. As a supervised method, it uses a Hidden Markov Model to rank sentences and a QR decomposition to produce the summary.

C-Attention (Li et al., 2017a): The cascaded attention based auto-encoder is proposed for estimating the relevancy of words and sentences.

GRU-GCN (Yasunaga et al., 2017): It is a supervised method that employs a Graph Convolutional Network on sentence graph. The sentence embedding obtained from a Recurrent Neural Network serves as the input node feature.

ParaFuse (Nayeem et al., 2018): MDS is formulated as multi-sentence compression. As the state-

[8]https://github.com/andersjo/pyrouge/tree/master/tools/ROUGE-1.5.5

of-the-art on DUC 2004, however, it needs some extra resource and toolkit, such as paraphrase bank and keyword extractor.

Best Review (Chu and Liu, 2019): A simple baseline selecting the best document to be summary based on word overlap.

Centroid (Rossiello et al., 2017): Word embeddings are exploited to boost the performance of centroid-based methods.

MeanSum (Chu and Liu, 2019): An end-to-end neural model is put forward to implement the abstractive summarization of business review documents. The summary is decoded from the mean of the representations of input reviews.

PG (See et al., 2017): It introduces a pointer-generator (PG) network that motivates the summarizer to copy original words from input via pointing, while preserving the ability to generate new words.

Hi-MAP (Fabbri et al., 2019): It proposes the integration of sentence-level MMR scores into the PG network in order to adapt the attention weights on a word-level. The MMR score is computed by the Maximal Marginal Relevance algorithm (Carbonell and Goldstein, 1998), which gives the goodness of the available sentence given already selected ones.

Our method Spectral: This is our spectral-based method specified in Alg. 2.1. According to different building strategies in §2.3, it has three versions: Spectral-tfisf, Spectral-ESE and Spectral-BERT.

Notice that on DUC 2004, the supervised methods are trained on datasets of earlier DUC evaluations or CNN and Daily Mail datasets (Hermann et al., 2015) according to their original papers. On Multi-News, they are both trained and tested on the dataset itself.

3.5 Main Results

Table 3.2 demonstrates the ROUGE results of various methods on the DUC dataset. The method ParaFuse is previously state-of-the-art on this dataset. From the table, our method Spectral-BERT outperforms ParaFuse by 1.2 percent in R-1 score and has a slightly lower R-2 and R-4 score (still ranking second). Notice that ParaFuse, as mentioned in §3.4, is not exactly a self-contained system. Compared with CLASSY04 (winner of the official evaluation), Spectral-BERT has an enormous advantage (say 3.7% and 2.7% higher R-1 and R-2 score, respectively). Notice that all supervised methods have a relatively low performance, since they are trained on datasets different from DUC 2004. This observation is consistent with the results reported in Fabbri et al. (2019).

Method		R-1	R-2	R-4
Supervised method	CLASSY04	0.376	0.090	1.51%
	GRU-GCN	0.385	0.095	1.32%
	Hi-MAP	0.358	0.089	1.46%
Unsupervised method	Lead	0.332	0.061	0.60%
	LexRank	0.360	0.075	0.82%
	C-Attention	0.391	0.093	1.61%
	ParaFuse	0.401	**0.120**	**1.87%**
	Spectral-tfisf	0.382	0.095	1.58%
	Spectral-ESE	0.404	0.108	1.67%
	Spectral-BERT	**0.413**	0.117	1.75%

Table 3.2: ROUGE results on the DUC 2004 dataset (our model is Spectral). The best ROUGE scores are highlighted in bold, and the second best are underlined.

Method	R-1	R-2	R-L
Lead	0.268	0.038	0.144
Centroid	0.246	0.029	0.138
Best review	0.280	0.035	0.153
MeanSum	0.289	0.037	0.159
Spectral-tfisf	0.283	0.036	0.147
Spectral-ESE	0.291	0.037	0.165
Spectral-BERT	**0.302**	**0.045**	**0.172**

Table 3.3: ROUGE results of unsupervised methods on the Yelp dataset (our model is Spectral). Best scores are in bold, and second best are underlined.

Table 3.3 and Table 3.4 show ROUGE results on Yelp and Multi-News, respectively. Our method Spectral-BERT has also beaten other unsupervised methods by a wide margin (say 1.3% higher R-L score than MeanSum and 2.1% higher R-SU4 score than C-Attention). Compared with the state-of-the-art supervised system on Multi-News (namely Hi-MAP), Spectral-BERT cannot rival its performance. However, Spectral-BERT has beaten the other supervised system (i.e. PG) according to R-2 and R-SU4 score.

We observe that a better matrix building strategy (stated in §2.3) has led to a considerable improvement of our method on all three datasets. Specifically, the BERT encoder brings about one percent improvement in R-2 score as compared with the tf-isf model. It proves that our method is flexible and can benefit from recent off-the-shelf pre-training techniques (Devlin et al., 2019; Clark et al., 2019).

Method		R-1	R-2	R-SU4
Supervised method	PG	0.419	0.129	0.165
	Hi-MAP	**0.435**	**0.149**	**0.174**
Unsupervised method	Lead	0.394	0.118	0.145
	LexRank	0.383	0.127	0.132
	C-Attention	0.386	0.125	0.146
	Spectral-tfisf	0.397	0.121	0.144
	Spectral-ESE	0.396	0.130	0.159
	Spectral-BERT	*0.409*	*0.136*	*0.167*

Table 3.4: ROUGE results on the Multi-News dataset (our model is Spectral). The best scores are in bold, the second best are underlined, the best among unsupervised methods are in *italics*.

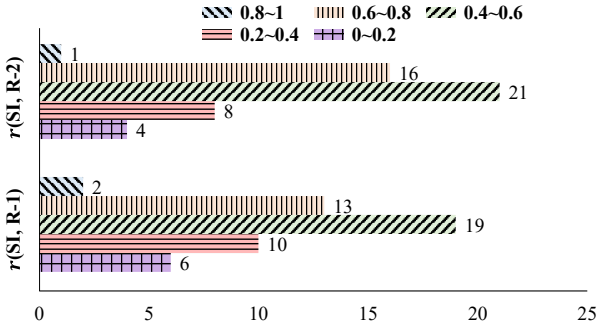

Figure 3.1: Pearson correlation coefficients r of spectral impact (SI) and ROUGE (R-1 or R-2) on DUC 2004. The sample size is the number of document clusters, i.e. 50. Best viewed in colors.

3.6 Linguistic Quality

To further assess the linguistic quality of different summaries, we employ Amazon Mechanical Turk[9] workers to judge the performance of three summarizers on a random sample of Multi-News (200 document clusters). A worker is asked to rate each summary on a scale of 1 (poor) to 5 (excellent) according to three criteria: **relevancy** (does the summary cover all the key information of document cluster?) and two criteria adopted by DUC 2005 evaluation (**non-redundancy** and **grammaticality**) (Dang, 2005). Table 3.5 shows the results. We observe that our method Spectral has the highest relevancy and non-redundancy score. Abstractive method C-Attention has a relatively low score of grammaticality. Notice that the non-redundancy scores of all summarizers are generally low, which shows that humans are more sensitive to the redundancy existing in the summary.

Method	Relev.	NonRed.	Gram.	Average
LexRank	4.19	2.74	**4.61**	3.85
C-Attention	4.32	3.18	3.25	3.58
Spectral-BERT	**4.57**	**3.32**	4.46	**4.12**

Table 3.5: Linguistic quality on Multi-News.

3.7 Hypothesis Validation

We provide the empirical evidence of our proposed spectral-based hypothesis. For each cluster of documents on DUC 2004, we construct a sample set of 500 summary candidates $\mathcal{S}$, each of which contains 3 original sentences selected randomly from the documents. The Pearson correlation coefficient (denoted by r) of spectral impact and the candidate goodness as a summary, when applied to the sample set, is computed and the derived histogram is shown in Figure 3.1. Each correlation coefficient falls into their corresponding bins. In the figure, our method Spectral-BERT and the ROUGE metrics are utilized because: (1) Spectral-BERT has been reported with a better empirical performance in §3.5; (2) The goodness of summary candidate in this scenario can be measured by the precision-oriented ROUGE scores, esp. R-1 and R-2, in that the word count of candidate $\mathcal{S}$ is varied in the sample. We note that there are no bins corresponding to negative correlation coefficients (r ranges from -1 to 1), and quite a few r's have a large score beyond 0.5 (the widely accepted threshold of a large r recommended by Cohen (2013)). This demonstrates that the two variables have a positive linear correlation, which supports our hypothesis. Similar results can be obtained when $\mathcal{S}$ contains a different number of sentences.

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

4 Related Work

Unsupervised MDS. There are a bunch of works working on different hypotheses and models in this field. PageRank alike algorithms (Erkan and Radev, 2004; Mei et al., 2010; Wang et al., 2017) utilize random walks with some redundancy avoiding measures. Regarding the document cluster as a manifold structure, (Wan et al., 2007; Cheng et al., 2011; Li et al., 2011) use the manifold ranking process on data. There are also quite a few neural architecture based models for a hidden semantic representation of sentences, documents or subtopics, such as (Ma et al., 2016; Li et al., 2017a,b; Zheng et al., 2019).
Spectral optimization. Optimizing eigen-related

metrics often leads to a specific collective optimization problem, which is believed to be hard in nature unless P = NP (Cook, 1971). Some particular examples (Van Mieghem et al., 2011; Chen et al., 2016) have been proved NP-hard. The typical solvers adopt the heuristics based on either perturbation theory (Chen et al., 2016) or QR decomposition (Li et al., 2015a; Chen et al., 2018).

5 Conclusion

We propose a novel hypothesis-driven method for unsupervised MDS, where the goodness of any summary candidate can be determined from a spectral perspective when dropping it from the document cluster. Various MDS tasks of different sizes and domains show a promising result of our method. Extending our method to an abstractive setting is meaningful future work.

Acknowledgments

We would like to thank all the reviewers for their helpful advice on various aspects of this work. This work is supported by National Natural Science Foundation of China (No. 61936012 and No. U19A2065) and Beijing Academy of Artificial Intelligence (BAAI).

References

Mohammed Al-Dhelaan. 2015. Starsum: A simple star graph for multi-document summarization. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 715–718.

Abdelghani Bellaachia and Mohammed Al-Dhelaan. 2014. Multi-document hyperedge-based ranking for text summarization. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1919–1922.

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

Chen Chen, Ruiyue Peng, Lei Ying, and Hanghang Tong. 2018. Network connectivity optimization: Fundamental limits and effective algorithms. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1167–1176.

Chen Chen, Hanghang Tong, B Aditya Prakash, Tina Eliassi-Rad, Michalis Faloutsos, and Christos Faloutsos. 2016. Eigen-optimization on large graphs by edge manipulation. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(4):49.

Zesheng Chen. 2018. Epidemic thresholds in networks: Impact of heterogeneous infection rates and recovery rates. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE.

Xue-Qi Cheng, Pan Du, Jiafeng Guo, Xiaofei Zhu, and Yixin Chen. 2011. Ranking on data manifold with sink points. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):177–191.

Eric Chu and Peter Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*, pages 1223–1232.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.

John M Conroy, Judith D Schlesinger, Jade Goldstein, and Dianne P O'leary. 2004. Left-brain/right-brain multi-document summarization. In *Proceedings of the Document Understanding Conference (DUC 2004)*.

Stephen A Cook. 1971. The complexity of theorem-proving procedures. In *Proceedings of the third annual ACM symposium on Theory of computing*, pages 151–158. ACM.

Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 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*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Kai Hong, John Conroy, Benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A repository of state of the art and competitive baseline summaries for generic news summarization. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1608–1616, Reykjavik, Iceland. European Language Resources Association (ELRA).

James Holland Jones. 2007. Notes on r0. *Stanford University*.

Cornelius Lanczos. 1950. *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*. United States Governm. Press Office Los Angeles, CA.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 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*, pages 4131–4141, Brussels, Belgium. Association for Computational Linguistics.

Cong Li, Huijuan Wang, and Piet Van Mieghem. 2012. Bounds for the spectral radius of a graph when nodes are removed. *Linear Algebra and its Applications*, 437(1):319–323.

Huiying Li, Yue Hu, Zeyuan Li, Xiaojun Wan, and Jianguo Xiao. 2011. Pkutm participation in tac2011. *Proceeding RTE*, 7.

Liangyue Li, Hanghang Tong, Yanghua Xiao, and Wei Fan. 2015a. Cheetah: fast graph kernel tracking on dynamic graphs. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 280–288. SIAM.

Piji Li, Lidong Bing, Wai Lam, Hang Li, and Yi Liao. 2015b. Reader-aware multi-document summarization via sparse coding. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Piji Li, Wai Lam, Lidong Bing, Weiwei Guo, and Hang Li. 2017a. Cascaded attention based unsupervised information distillation for compressive summarization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2081–2090, Copenhagen, Denmark. Association for Computational Linguistics.

Piji Li, Zihao Wang, Wai Lam, Zhaochun Ren, and Lidong Bing. 2017b. Salience estimation via variational auto-encoders for multi-document summarization. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Hui Lin and Jeff Bilmes. 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*, pages 510–520, Portland, Oregon, USA. Association for Computational Linguistics.

He Liu, Hongliang Yu, and Zhi-Hong Deng. 2015. Multi-document summarization based on two-level sparse representation model. In *Twenty-ninth AAAI conference on artificial intelligence*.

Ying Liu, Albert A Gayle, Annelies Wilder-Smith, and Joacim Rocklöv. 2020. The reproductive number of covid-19 is higher compared to sars coronavirus. *Journal of travel medicine*.

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

Qiaozhu Mei, Jian Guo, and Dragomir Radev. 2010. Divrank: the interplay of prestige and diversity in information networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1009–1018. Acm.

Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. 2018. Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1191–1204, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9, Montréal, Canada. Association for Computational Linguistics.

B Aditya Prakash, Deepayan Chakrabarti, Nicholas C Valler, Michalis Faloutsos, and Christos Faloutsos. 2012. Threshold conditions for arbitrary cascade models on arbitrary networks. *Knowledge and information systems*, 33(3):549–575.

Gaetano Rossiello, Pierpaolo Basile, and Giovanni Semeraro. 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, pages 12–21, Valencia, Spain. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Dragan Stevanovic. 2014. *Spectral radius of graphs*. Academic Press.

G. W. Stewart. 1990. Matrix perturbation theory.

Piet Van Mieghem, Dragan Stevanović, Fernando Kuipers, Cong Li, Ruud Van De Bovenkamp, Daijie Liu, and Huijuan Wang. 2011. Decreasing the spectral radius of a graph by link removals. *Physical Review E*, 84(1):016101.

Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. 2007. Manifold-ranking based topic-focused multi-document summarization. In *IJCAI*, volume 7, pages 2903–2908.

Kexiang Wang, Tianyu Liu, Zhifang Sui, and Baobao Chang. 2017. Affinity-preserving random walk for multi-document summarization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 210–220, Copenhagen, Denmark. Association for Computational Linguistics.

Yang Wang, Deepayan Chakrabarti, Chenxi Wang, and Christos Faloutsos. 2003. Epidemic spreading in real networks: An eigenvalue viewpoint. In *22nd International Symposium on Reliable Distributed Systems, 2003. Proceedings.*, pages 25–34. IEEE.

DJA Welsh. 1976. Matroid theory. 1976.

Kang Yang, Kamal Al-Sabahi, Yanmin Xiang, and Zuping Zhang. 2018. An integrated graph model for document summarization. *Information*, 9(9):232.

Kang Yang, Hongye He, Kamal Al-Sabahi, and Zuping Zhang. 2019. Ecforest: Extractive document summarization through enhanced sentence embedding and cascade forest. *Concurrency and Computation: Practice and Experience*, 31(17):e5206.

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

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 452–462, Vancouver, Canada. Association for Computational Linguistics.

Xin Zheng, Aixin Sun, Jing Li, and Karthik Muthuswamy. 2019. Subtopic-driven multi-document summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3153–3162, Hong Kong, China. Association for Computational Linguistics.