

Multi-Document Summarization Based on Locally Relevant Sentences

Esaú Villatoro-Tello*, Luis Villaseñor-Pineda*, Manuel Montes-y-Gómez* and David Pinto-Avendaño†

*Laboratory of Language Technologies, Department of Computational Sciences,
National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico.

Email: {villatoroe, villasen, mmontesg} @ccc.inaoep.mx

†Faculty of Computer Science,
University of Puebla (BUAP), Mexico.

Email: dpinto@solarium.cs.buap.mx

Abstract—Multi-document summarization systems must be able to draw the “best” information from a set of documents. In this paper we propose a novel extractive approach for multi-document summarization based on the detection of *locally relevant sentences*. Our main hypothesis is that by extracting relevant sentences from each document within a collection, instead of considering all documents at once, the final multi-document summary will be of higher quality. Performed experiments showed that the proposed method is able to outperform conventional baselines as well as traditional approaches by constructing summaries of high quality according to the ROUGE evaluation metrics.

I. INTRODUCTION

Multi-document summarization aims to produce a summary delivering the majority of information content from a collection of topic related documents [1], [2]. Multi-document summarization involves multiple sources of information that overlap and supplement each other, and being contradictory at occasions. Therefore, the key tasks are not only identifying and extracting redundant information across documents, but also recognizing novelty and ensuring that the final summary is both coherent and complete.

Automatic multi-document summarization has attracted much attention in recent years both in the research community and business community since it exhibits the practicability in document management and search systems. A multi-document summary can be used to concisely describe the information contained in a cluster of documents and facilitate the users to understand the main topic within the document cluster

In this paper we propose a novel extractive approach for multi-document summarization based on the detection of *locally relevant sentences*. Our method is divided in two major steps: *i)* first, we treat each document individually, in order to detect relevant sentences for each of them, i.e., *locally relevant sentences*. The output of this step can be seen as a set of individual extractive summaries. Then, *ii)* in the second step, we focus on finding all the common and different themes across the individual generated summaries, so we can finally select and extract most representative elements to create our final extractive multi-document summary.

The rest of the paper is organized as follows. Section II discusses some related work. Section III describes the

proposed method. Section IV presents the experimental results. Finally, section V depicts our conclusions and future work.

II. RELATED WORK

Traditionally, multi-document summarization has been seen as a two steps problem: 1) identify common and different information (i.e., *themes*) among the document set; 2) select the most representative elements contained in each *theme*, which will be included in the final summary. Additionally, some research groups consider a third step, known as a *generation* step [3], [4], [5], which consists in merge and reformulate a new grammatical text based on extracted elements.

According to these intuitive ideas, multi-document summarization has been treated through the combination of clustering and sentence ranking strategies. In [4] and [5], they start by identifying *themes* among the set of documents, i.e., sets of similar text units (paragraphs), for this they employed a *clustering* strategy based on decision rules to determine when two text units are similar or when are not. To compute similarities between text units, these are mapped to vectors of features that include single words, noun phrases, proper nouns, and synsets from the Wordnet. Then, an *information fusion* algorithm is employed to select elements that should be included in the final summary.

In [6] themes detection is made by an agglomerative clustering algorithm that operates over the TF-IDF vector representations of the documents. At the contrary of above systems, documents are modeled as bags-of-words. In a second stage, cluster centroids are used to identify most representative sentences in each cluster. At the end, extracted sentences are ordered chronologically and given as final summary to the user.

Recent works [2], [7] are simple extensions or variations of generic multi-document summarization. In [2] all sentences within documents set are ordered by a ranking algorithm, once all sentences have been ranked, the highest ranked sentences are given as final summary to the user. Whereas in [7] sentences are ranked by its centroid value, then, a trimming process based on manual generated rules is applied in order to generate a more coherent summary.

A common problem with these “generic” techniques is that they are vulnerable to include irrelevant information in the final summary. This happens since all sentences from all documents within the set are considered for the *theme* identification (i.e., clustering) stage. Hence, the information of certain documents could be more representative than others just by being longer. This fact may result in a bad clustering because formed clusters could be either a few with low cohesion or many with high cohesion. Whatever is the case, the final summary is always affected.

Our proposed method intends to solve this problem by first identifying relevant information contained in each document within the set, i.e., *local relevant sentences*. Our main hypothesis is that, by doing this, final multi-document summaries will be of higher quality since only relevant information will be considered during the theme identification stage.

III. PROPOSED METHOD

The proposed method consists of two main modules, which are:

- i *Extraction of relevant sentences*, where the main goal is to identify and extract the most relevant sentences from each document within the given document set. The output of this module can be seen as set of individual extractive summaries.
- ii *Identification of themes*, which goal is to find all common and different information across previously constructed individual summaries, generating clusters of individual summaries. Afterwards, a selective process is performed, its goal is to select and extract the most representative elements from each theme to construct the final multi-document summary.

It is worth mentioning that the second step in our method, i.e., themes identification and most representative elements selection, corresponds to the traditional ideas applied to solve the problem of multi-document summarization. Our main contribution is the idea of including, previous to these steps, a method that treats each document individually, and selects only relevant information from each document.

A. System Implementation

Figure 1 gives a general overview of our multi-document summarization method.

First step: extraction of locally relevant sentences. The core module of our multi-document summarizer relies in this first step. Here, we focus on the creation of single-document summaries by the selection of relevant sentences from each input text. Particularly we used the method proposed in [8]. This is a supervised method for single document summarization, where documents are represented through word-based features (i.e., n -grams)¹.

Traditional methods for supervised text summarization use “heuristically motivated” features to represent the sentences

¹At this point, any other single document summarization method could be used.

[9], [10]. Nevertheless, they have the major disadvantage of being highly related to a target domain. On the contrary to these works, our single document summarizer considers word-based features in order to increase the summarization flexibility by lessening the domain and language dependency. In particular, we use n -grams (sequences of n consecutive words) as sentence features. Thus, in our model each sentence is represented by a feature vector that contains one boolean attribute for each n -gram that occurs in the training collection.

We choose this representation, since as is established in [8], n -grams features are adequate for fine-grained classification task such as text summarization. For performed experiments, we only consider sequences up to three words, i.e., from 1-grams to 3-grams.

As main classifier we employed the Naïve Bayes strategy [9], which has proved to be quite competitive for most text processing tasks including text summarization. It basically computes for each sentence s its probability (i.e., a score) of been included in a summary S given the k features $F_j; j = 1..k$. This probability can be expressed using Bayes’ rule as follows:

$$P(s \in S | F_1, F_2, \dots, F_k) = \frac{P(F_1, F_2, \dots, F_k | s \in S) P(s \in S)}{P(F_1, F_2, \dots, F_k)} \quad (1)$$

Assuming statistical independence of the features:

$$P(s \in S | F_1, F_2, \dots, F_k) = \frac{\prod_{j=1}^k P(F_j | s \in S) P(s \in S)}{\prod_{j=1}^k P(F_j)} \quad (2)$$

where $P(s \in S)$ is a constant and $P(F_j | s \in S)$ and $P(F_j)$ can be estimated directly from the training set by counting occurrences.

Second step: themes identification. This step considers traditional approach to solve the problem of multi-document summarization. For this process we used the *Star Clustering Algorithm* [11]. Advantages of this algorithm are: 1) it induces in a natural form the number of clusters, and 2) it finds the natural topic structure of the documents’ space.

Star Clustering Algorithm is based on a greedy cover of a thresholded similarity graph G_σ giving as output star shaped sub-graphs, where the central element from each star it’s called *center*, and adjacent elements to the center are called *satellites*. A *correct star cover* is defined as a star cover that assigns the types “center” and “satellite” in such a way that: (1) a star center is not adjacent to any other star center and (2) every satellite vertex is adjacent to at least one center vertex of equal or higher degree.

For our experiments, first, we computed similarities among input documents² using the well known cosine formula [12]. Whereas for the construction of G_σ three different values for σ were considered, which take into account the statistical information from the documents’ similarity matrix: (1) $\sigma = \bar{x}$,

²Input documents are in fact individual extractive summaries (see Fig. 1)

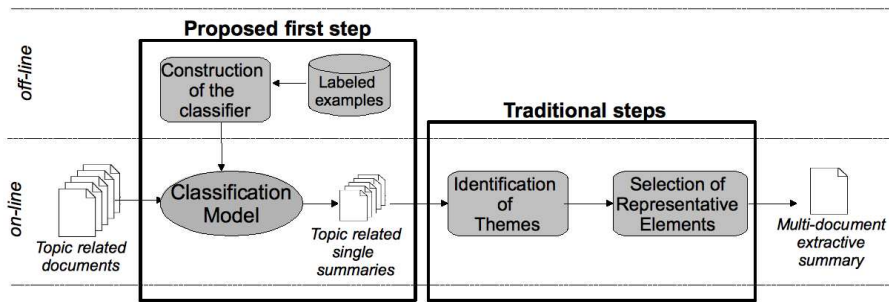


Fig. 1. General overview of the proposed method

TABLE I
DUC 2003 (TASK 4) DATA SETS

Name	Language	Domain	Num of Collections	Number of sentences	Number of relevant sentences
DUC-03	English	News	30	33666	1995
task-4		reports	(22 docs x collection)		(5.99%)

(2) $\sigma = \bar{x} + \delta$ and (3) $\sigma = \bar{x} - \delta$; where \bar{x} represents the statistical mean, and δ the standard deviation, both computed from the similarity matrix. By doing this, we intent to make our method adaptable and adequate to the nature of the document collection. Furthermore, we avoid the user intervention in the process of defining an appropriate threshold.

Finally, our methodology to select the most representative elements was: from the biggest to the smallest *star cluster*, we review and extract from each “center vertex” the first sentence, that has not been previously extracted, until the target summary size is reached.

IV. EXPERIMENTS AND RESULTS

A. Data Set

For our experiments we employed DUC³-2003 data set, particularly, we used the task 4 corpus of DUC-2003. This is a labeled corpus, i.e., each sentence in each document has a label that indicates whether the sentence is “relevant” or “not relevant”. Along with each document collection, a set of four human generated summaries are given, which will be employed for evaluation purposes. Table I shows some statistics about the DUC-2003 data set.

As can be seen in Table I, the distribution of classes is highly unbalanced since only 5.92% of sentences are relevant instances. For our experiments, we arbitrary select 80% of the collections for training (i.e., 22 collections), and for the test phase we used the remaining 20% of the collections (i.e., 8 collections).

B. Evaluating the Classification Model

As we have mention before, our first stage is single-document summarizer based on a machine learning approach.

Hence, we can measure its performance in terms of *precision* (p), *recall* (r), and *accuracy* (a).

Table II shows the number of computed n -grams (from 1 to 3-grams) for the set of training collections. In order to reduce the dimensionality of the features space, we applied the well known *Information Gain* algorithm [13] to identify the most representative features (See table II).

TABLE II
COMPUTED FEATURES (n -GRAMS) IN THE DUC-2003 CORPUS

	Original			Selected	
	1-grams	2-grams	3-grams	Total	Total
DUC-2003-80%	25937	205052	339696	570685	1291

Table III show results obtained for our classifier. First row (DUC-2003-80%) show results obtained over the training set considering as a evaluation strategy a *10-fold cross validation*. Whereas second row (DUC-2003-20%) show results obtained evaluating over the test set.

TABLE III
SINGLE-DOCUMENT SUMMARIZER PERFORMANCE

	Features					
	Single words			n -grams		
	precision	recall	accuracy	precision	recall	accuracy
DUC-2003-80%	93.04	97.35	91.03	93.57	97.77	91.91
DUC-2003-20%	95.94	92.28	89.60	96.35	92.65	89.62

As we can see, n -grams representation allows obtaining a higher performance than using just *single words* as features. For the case of DUC-2003-80% the differences between the 97.35% of recall using single words, and 97.77% using n -grams, implies a profit of 104 correctly classified instances. Notice that in this first stage, the generated single-document summary has no restrictions of size, i.e., we preserve all the sentences that the classifier selected as relevant; is until the second stage where the size restriction is considered.

It is worth pointing out that this intermediate evaluation (i.e., the classifier evaluation in table III) do not represent our main results. It is in the following section (IV-C) where our multi-document summarizer is evaluated.

³Document Understanding Conference (<http://duc.nist.gov>)

$$ROUGE - N = \frac{\sum_{S_i \in \{ReferenceSummary\}} \sum_{gram_n \in S_i} Count_{match}(gram_n)}{\sum_{S_i \in \{ReferenceSummary\}} \sum_{gram_n \in S_i} Count(gram_n)} \quad (1)$$

C. Evaluating Multi-document Summaries

We used the ROUGE [14] toolkit for evaluation, which was adopted by DUC for automatically summarization evaluation. It measures summary quality by counting overlapping units such as the n -gram, word sequences and word pair between the candidate summary (automatically generated) and the reference summary (human generated). ROUGE-N: is an n -gram recall measure computed as indicated in the formula 1.

Where S_i refers to sentence i within the reference summary, n stands for the length of the n -gram, $gram_n$, and $Count_{match}(gram_n)$ is the maximum number of n -grams co-occurring in a candidate summary and a set of reference summaries. $Count(gram_n)$ is the number of n -grams in the reference summaries.

The ROUGE toolkit reports separate scores for 1, 2, 3, and 4-grams, and also for longest common subsequence occurrences. Among these different scores, uni-gram-based ROUGE score (ROUGE-1) has been shown to agree with human judgment most [14]. We show five of the ROUGE metrics in the experimental results, at a confidence level of 95%: ROUGE-1 to ROUGE-4 and ROUGE-L (based on longest common subsequence).

D. Baseline Configuration

Two different methods for computing a baseline have been defined by the scientific community from the DUC conference:

- 1) Most recent document: It consist on selecting the first N lines (or bytes) from the most recent document in the entire collection. From here we call this baseline 1.
- 2) First N lines: It consist on selecting from each document within the collection the first N lines (or bytes). Usually only one sentence from each document is selected. From here we call this baseline 2.

E. Results

Table IV and V shows result achieved for our multi-document summarizer method when we compare our generated summaries against one reference summary and against four reference summaries respectively. Each row indicates in the configuration column the parameters employed for our Multi-Document Summarization System (MDSS) to generate its corresponding results. As we mention before, we used the cosine formula to compute document similarities, and we define three ways of selecting σ for the graph construction. For all experiments, summary target size was set to 200 words.

We can observe that in both tables (Table IV and V), that the proposed method allow to generate more pertinent summaries than the leading baselines. From these tables we can say that our system allow to generate summaries that have almost

TABLE IV
PROPOSED METHOD AGAINST ONE REFERENCE SUMMARY

Configuration	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L
MDSS-COS (\bar{x})	0.39057	0.06445	0.01966	0.00879	0.34492
MDSS-COS ($\bar{x} - \delta$)	0.38004	0.05777	0.01477	0.00626	0.33579
MDSS-COS ($\bar{x} + \delta$)	0.40482	0.08131	0.02922	0.01585	0.36435
Baseline 1	0.25111	0.04065	0.01594	0.0078	0.22983
Baseline 2	0.25322	0.0265	0.00776	0.00434	0.2372

TABLE V
PROPOSED METHOD AGAINST FOUR REFERENCE SUMMARIES

Configuration	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L
MDSS-COS (\bar{x})	0.39038	0.06521	0.0212	0.00921	0.34464
MDSS-COS ($\bar{x} - \delta$)	0.37539	0.05486	0.0142	0.00568	0.33281
MDSS-COS ($\bar{x} + \delta$)	0.39223	0.07562	0.02659	0.01447	0.3521
Baseline 1	0.25237	0.03879	0.01507	0.0074	0.23095
Baseline 2	0.26084	0.02808	0.00868	0.00418	0.24514

the same agreement degree with one than with four human generated summaries.

Also, we can observe that higher performance values are obtained when a *hard threshold* (i.e., $\bar{x} + \delta$) is employed for the generation of the star shaped clusters. Setting our clustering algorithm with a *hard threshold* means that encountered themes will be more “hardly related”, i.e., a major number of “hardly reliables” clusters is generated.

As mention in [11], it is possible for the star clustering algorithm to have more than one vertex as a possible *star center* element. When this situation occurs, only one is arbitrarily selected to be the *star center*. This situation frequently occurs when a “relaxed threshold” (i.e., $\bar{x} - \delta$) is employed for the generation of clusters, resulting in a less number of clusters with unreliable *star centers*. When this occurs, we are not certain if selected elements are correctly representing its group; this is possible to confirm in Tables IV and V since this relaxed configuration is the one that obtains lower results.

Additionally, we performed a second experiment where only the traditional approach (see Figure 1) is considered, i.e., we do not consider our proposed first step in the process of generating the multi-document summary. The goal was to probe that by eliminating in a first step all irrelevant information contained in each document, it is possible to construct higher quality summaries.

Table VI show results obtained when applying only a traditional approach considering three different thresholds. Each row indicates in the configuration column the parameters employed for the star clustering algorithm (STAR) to generate its corresponding results. As we can see, the proposed method is able to generate more pertinent summaries than the traditional approach. These results support the importance of detecting locally relevant sentences contained in each document before the themes identification phase.

TABLE VI

TRADITIONAL APPROACH VS THE PROPOSED METHOD (CONSIDERING FOUR REFERENCE SUMMARIES)

Configuration	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L
STAR-COS (\bar{x})	0.35986	0.04979	0.01392	0.00431	0.31933
STAR-COS ($\bar{x} - \delta$)	0.33646	0.04373	0.00862	0.0038	0.2963
STAR-COS ($\bar{x} + \delta$)	0.24189	0.04181	0.01403	0.00663	0.2169
Proposed Method	0.39223	0.07562	0.02659	0.01447	0.3521
Baseline 1	0.25237	0.03879	0.01507	0.0074	0.23095
Baseline 2	0.26084	0.02808	0.00868	0.00418	0.24514

V. CONCLUSION

This paper proposed a novel method for multi-document extractive summarization. The proposed method allows constructing multi-document summaries based on high quality clusters generated from individual document summaries. For this, in a first supervised stage, we detect and extract *locally relevant sentences* contained in each document within the set. By doing this, we guarantee the selection of only true important content from each document. Then, in a second unsupervised stage, the star clustering algorithm finds all themes across documents' summaries. At the end, only most representative elements are selected for the construction of the final summary.

The main contribution of this paper is that it represents the first attempt for generating multi-documents summaries by treating documents individually instead of considering all the documents as one, which is the basic idea of traditional approaches.

Our performed experiments showed that the proposed method is able to construct more pertinent summaries according to the ROUGE measures. Obtained results outperform both conventional baselines and also the traditional scheme, motivating us to keep working in this field. As future work we are planning a major study evaluating our method on a more recent data sets to provide more evidence of the efficiency of the proposed method, as well as considering different summary's size (e.g., 100 and 400 words summaries).

ACKNOWLEDGMENTS

This work was done under partial support of CONACyT (scholarship 165545, and project grants 83459 and 82050).

REFERENCES

- [1] D. R. Radev, E. Hovy, and K. McKeown, "Introduction to the special issue on summarization," *Computational Linguistics*, vol. 28, no. 4, pp. 399–408, 2002.
- [2] X. Wan, J. Yang, and J. Xiao, "Manifold-ranking based topic-focused multi-document summarization," in *IJCAI '07*, 2007, pp. 2903–2908.
- [3] K. R. McKeown and D. R. Radev, "Generating summaries of multiple news articles," in *Proceedings of SIGIR '95*, Seattle, Washington, 1995, pp. 74–82.
- [4] K. R. McKeown, J. Klavans, V. Hatzivassiloglou, R. Barzilay, and E. Eskin, "Towards multidocument summarization by reformulation: progress and prospects," in *AAAI/IAAI*, 1999, pp. 453–460.
- [5] R. Barzilay, K. R. McKeown, and M. Elhadad, "Information fusion in the context of multi-document summarization," in *Proceedings of ACL '99*, 1999.
- [6] D. R. Radev, H. Jing, M. Stys, and D. Tam, "Centroid-based summarization of multiple documents," *Information Processing and Management*, vol. 40, pp. 919–938, 2004.

- [7] K. Sarkar, "Improving multi-document text summarization performance using local and global trimming," in *Proceedings of the First international Conference on Intelligent Human Computer Interaction*, 2009, pp. 272–282.
- [8] E. Villatoro-Tello, L. Villaseñor-Pineda, and M. Montes-y-Gómez, "Using word sequences for text summarization," in *Proceedings of the 9th International Conference on Text, Speech and Dialogue (TSD 2006)*, Brno, Czech Republic, 2006, pp. 293–300.
- [9] J. Kupiec, J. O. Pederson, and F. Chen, "A trainable document summarizer," in *Proceedings of the 18th ACM-SIGIR Conference on Research and Development in Information Retrieval*, Seattle, USA, 1995, pp. 68–73.
- [10] T. W. Chuang and J. Yang, "Text summarization by sentence segment extraction using machine learning algorithms," in *Proceedings of the ACL'04 Workshop*, Barcelona, Spain, 2004.
- [11] J. Aslam, K. Pelehov, and D. Rus, "A practical clustering algorithm for static and dynamic information organization," in *Proceedings of the 1999 Symposium on Discrete Algorithms*, 1999, pp. 194–218.
- [12] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison Wesley, 1999.
- [13] T. Mitchell, *Machine Learning*. McGraw Hill, 1997.
- [14] C. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out: Proceedings of the ACL'04 Workshop*, M. Marie-Francine and S. S., Eds., Barcelona, Spain, 2004, pp. 74–81.