Towards Coherent Single-Document Summarization: An Integer Linear Programming-based Approach

Rodrigo Garcia
Federal Rural University of Pernambuco
Recife, PE
carlos.rodrigogarcia@ufrpe.br

Rinaldo Lima
Federal Rural University of Pernambuco
Recife, PE
rinaldo.jose@ufrpe.br

Bernard Espinasse
Aix-Marseille University - LSIS UMR CNRS
Marseille, France
bernard.espinasse@lsis.org

Hilário Oliveira
Federal University of Pernambuco
Recife, Brazil
htao@cin.ufpe.br

ABSTRACT

Automatic Text Summarization (ATS) is a viable option to reduce the content of textual documents, e.g., as a possible preprocessing step in many text mining applications. Single-document extractive summarizers have been developed based on different approaches, but many of them have the drawback of producing summaries with low coherence among the selected sentences in the generated summaries. In this paper, we present an unsupervised summarization system as an attempt towards coherent extractive single-document summarization. This system relies on Integer Linear Programming (ILP) as an optimization technique for selecting the smallest subset of sentences of a document maximizing the coverage of relevant concepts. Furthermore, our solution uses a graph-based algorithm for two goals: representing both sentences and concepts and enabling local coherence scoring among the sentences in the generated summaries. The proposed system is evaluated on two single-document benchmark datasets (DUC 2001-2002) using ROUGE measures, and compared with other state-of-the-art summarizers. The achieved results are very competitive.

CCS CONCEPTS

•**Information systems** → **Summarization;** *Digital libraries and archives;* Information extraction; •**Computing methodologies** → Semantic networks;

KEYWORDS

Single-document Summarization, Extractive Summarization, Coherence, Entity Graph, Integer Linear Programming

ACM Reference format:
Rodrigo Garcia, Rinaldo Lima, Bernard Espinasse, and Hilário Oliveira. 2018. Towards Coherent Single-Document Summarization: An Integer Linear Programming-based Approach. In *Proceedings of ACM SAC Conference, Pau,France, April 9-13, 2018 (SAC'18),* 8 pages.
DOI: https://doi.org/10.1145/3167132.3167211

SAC'18, Pau,France
 978-1-4503-5191-1/18/04…$15.00
DOI: https://doi.org/10.1145/3167132.3167211

1 INTRODUCTION

Due to the constant expansion of the Internet, the amount of data and textual documents has grown exponentially in recent years. Despite the constant improvement in the development of web search engines, identifying useful information from this huge amount of data is impractical if performed manually. In this context, Automatic Text Summarization (ATS) can be very useful for creating a summary from a document (single-document approach) or a collection of documents (multi-document approach) containing the portion of the input document(s) that only conveys the most important and relevant information from the original document(s).

ATS systems can be classified into two major approaches: *extractive* and *abstractive*. The former generates summaries by selecting the most salient (relevant) sentences from the input document and concatenating them to form the final summary [6, 19]; whereas the latter tries to produce more like-human summaries by transforming the original text, paraphrasing it [2].

The extractive summarization approach has been the most studied so far. However, most current extractive summarization systems usually produce summaries with two major issues: loose sentences lacking relationship among them, and dangling coreferences that breaks the natural discourse flow. Currently, adding coherence in extractive summarization is one of the main open problems in ATS. The present work aims at mitigating this limitation.

In this paper we present an extractive single-document summarization system that selects the smallest possible set of sentences maximizing the coverage of relevant concepts from the input document, taking into account the maximum size of the summary that should be generated. In other words, it is based on an optimization procedure that cast the summarization task as an optimization problem, as previously done in [6, 9, 22]. However, contrary to the aforementioned works, the proposed system relies on a specific graph-based representation of entity occurrences in a document as a means for estimating coherence scores for sentences. This representation is implemented as a bipartite graph which enables the coherence score be calculated as the outdegree of the nodes which represent sentences. This idea was first introduced in [10] and has demonstrated encouraging performance results in summarization and related tasks [19, 20, 23].

The rest of this paper is structured as follow: Section 2 briefly described related work on single-document summarization based on

ILP. Section 3 presents our proposal of an extractive summarization system based on optimization using ILP and entity graph-based representation to model coherence of sentences in a document. Section 4 discusses experimental results on two reference dataset, and compares these results with other systems. Finally, Section 5 concludes this paper and outlines future work.

2 ILP AND EXTRACTIVE SUMMARIZATION

Recently, several extractive summarizers have adopted the strategy to model the summarization process as a constrained maximum coverage problem, i.e., maximize essential aspects of the summary, e.g., relevance and coherence, while taking into account certain restrictions, such as the maximum size of the summary. Their assumption is that a good summary can be formed by selecting sentences containing as many important concepts as possible from the document [9]. These systems are mainly rooted on an ILP-based solver which provides constrained optimization solutions according to a cost function and constraints linear in a set of integer variables. Although global inference is actually a NP-hard problem, some approximate solutions using ILP have been reported in the literature for ATS [15].

A concept-based ILP model for summarization is proposed by Gillick et al. (2009) as a strategy that casts sentence selection as a maximum coverage problem. The assumption of the model is that the value of a summary is defined as the sum of the weights of the unique concepts it contains. Thus, a summary only benefits from including each concept once.

Boudin et al. (2015) extended the model to reduce the number of concepts in the model. It uses a concept pruning technique, and the authors have shown empirical evidence that concept pruning can lead to multiple optimal solutions instead of just one.

In Schluter and Sogaard (2015), contrary to the previous works, they evaluate the maximum coverage objective for extractive summarization considering syntactic structures and semantic concepts. More precisely, they replace bigram concepts with new ones based on syntactic dependencies, semantic frames, as well as named entities. Furthermore, they show that using such concepts can lead to significant improvements in performance outside of the newswire domain.

Parveen et al. (2015) proposed an extractive graph-based unsupervised technique for summarizing single documents optimizing three important properties of summarization, i.e. importance, non-redundancy and local coherence. The input document is represented by a bipartite graph whose nodes are sentences and entities. A graph-based ranking algorithm, the Hyperlink-Induced Topic Search, is applied on this graph for computing the rank of sentences based on their relevance. Non-redundant and locally coherent summaries are produced through an optimization process using ILP.

Except for the Parveen's work, all the aforementioned works do not take into account the coherence of the generate summaries. This is a major drawback because it can lead to summaries lacking discourse flow. Among the systems studied above, [6] and [22] both use ILP only for maximizing informativeness (relevance), while [9] has employed ILP to deal with redundancy issues.

In order to alleviate such limitations, this work proposes an unsupervised ILP-based approach to single-document summarization that takes into account both the importance and the local coherence of the sentences in the produced summaries. On the one hand, it tries to maximize the informativeness of the concepts in the summary. On the other hand, it takes into account the coherence of the generated summary aiming at generating the best possible readable summary. The proposed system is in part similar to the Parveen's system which uses ILP for maximizing informativeness and coherence, but it represents document at concept level, instead of Parveen's work which deals with documents at sentence level.

3 OUR METHOD

This section presents the proposed summarizer that not only maximizes informativeness of the summaries, but takes into account local coherence of the sentences.

3.1 Optimization Process for Summarization

The proposed summarizer is based on an optimization procedure that cast the summarization task as an optimization problem, as previously done in [9] [6] [22]. In other words, it selects the smallest possible number of sentences that maximize coverage of relevant concepts from the input document, taking into account the maximum size of the summary to be generated. In addition, it relies on a specific graph-based representation of entity occurrences in a document as a means for estimating coherence scores for the sentences. This representation is implemented as a bipartite graph which contains nodes representing sentences and entities. This same graph is used as input for calculating local coherence scores as a centrality measure based on the nodes representing sentences. This idea was first introduced by [10] and has shown encouraging performance results in summarization and related tasks [19] [20] [23].

An overview of the main components of the proposed summarizer is depicted in Fig 1. These components are described in detail next.

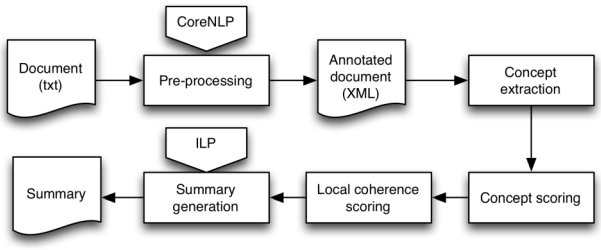

Figure 1: Overview of our summarization method.

3.2 Preprocessing

Text preprocessing is performed to identify the concepts in the input documents. We rely on the Stanford CoreNLP Toolkit [14] for annotating the documents by performing the following natural language processing subtasks: tokenization, sentence splitting, Part-of-Speech (POS) tagging, lemmatization, dependency parsing, and coreference resolution. In addition, we remove all stopwords from the sentences. Coreference resolution is employed to find anaphoric

expressions referring to the same entity or concept in a document. All personal pronouns, except for the pronoun 'it', are replaced by the referring entities. At the end of this preprocessing step, the annotations are persisted in XML documents.

3.3 Concept Extraction

In this step, unigram nouns are extracted as concepts, e.g., in the sentence *"Paul wrote a book on Artificial Intelligence"*, the following concepts will be extracted "Paul", "book", "Artificial", and "Intelligence". This reflects the intuition that such terms are important for text summarization because they likely describe real world entities mentioned in the document.

3.4 Concept Scoring

A wide variety of methods for weighing the importance of concepts have been proposed in the literature. Among them, *Term Frequency* (TF), *Normalized Term Frequency* (NTF), and *Term Frequency/Inverse Sentence Frequency* (TF-ISF) were adopted in this work.

TF is simply the number of occurrences of a term (word) in a document d, i.e., $TF(w)$; while the NTF of a word w in d is calculated by Equation 1:

$$NTF(w, d) = \frac{TF(w)}{n} \tag{1}$$

where: n is the number of sentences in document d. These two scoring methods are based on the assumption that the higher the frequency of a term, the more important it is to the document.

The TF-ISF [5] is derived from the classical Information Retrieval scoring method Term Frequency/Inverse Document Frequency (TF-IDF). The main difference between them resides in the granularity level when computing the second term of their formula: IDF is computed at the document-level, reflecting the distribution of terms within the corpus [21]; while ISF is computed at sentence-level, i.e., reflecting the distribution of terms with respect to all the sentences for a document [5].

The TF-ISF score of a word w is computed by Equation 2

$$TF - IFS(w) = TF(w) \times \log(\frac{n}{TS(w)}) \tag{2}$$

where: TF(w) as defined above, n is the total number of sentences in the document, and TS(w) is the total number of sentences in which w occurs. This scoring method is based on the assumption that if a word is frequent and it appears in a few sentences of a document, then it should be included into the summary. The choice for the above three scoring methods was motivated by the fact that they were among the best scoring methods assessed in [7].

3.5 Coherence Scoring

In this work, the method for both representing the sentences and modeling local coherence is based on the Entity Graph (EG) introduced by Guineaudau and Strube (2013). The EG is used for local coherence modeling that captures the distribution of discourse entities across sentences. In our solution, the local coherence scoring is performed in two steps:

(1) first, the document is represented by a bipartite graph, a.k.a Entity Graph (EG), containing two distinct sets of nodes: one corresponding to sentences, and the other denoting entities (Fig. 2). Concepts or entities are defined by unigrams identified as nouns and personal pronouns at the preprocessing step. In addition, syntactic information gathered in the preprocessing phase also provides additional information for characterizing an edge as follows: an entity in the syntactic role of subject (S), object (O), or other(X) obtains weights 3, 2, and 1, respectively.

(2) then, the EG is transformed into a new directed graph, the Projection Graph (PG), containing only nodes corresponding to sentences connected by an edge if they share at least one common entity (Fig. 3). Differently from the EG, the PG is directed and follows the sentence order in the document. This PG representation is used for calculating the local coherence of a document relying only on centrality measures applied to the nodes of the PG. More precisely, the local coherence score for a given sentence *si* of the PG shown in Fig. 3 is equal to the outdegree score given by Equation 3.

$$coherence(s_i) = Outdegree(s_i, PG) \tag{3}$$

Equation 3 calculates, for every sentence from the projection graph *PG*, the total weight of all the edges leaving the sentence (centrality measure). Thus, the higher is the centrality value for a sentence S in PG, the more connected S is with other sentences. This centrality measure has the advantage of assessing the importance of sentences by taking into account the stronger entity transitions instead of simply assigning scores based on the most frequent words.

The Entity Graph model was adopted in this work due to two main reasons: it provides an easy way to assess to which extent a sentence is connected (in terms of discourse entities), with other sentences in a document [10]; it has achieved good performance in recent work on summarization [18, 19] and other related tasks [20, 23].

In what follows, we illustrate the coherence scoring process. Consider the document composed by the following sentences:

S_1: $Hurricane_{[E_1]}$ $Gilbert_{[E_2]}$ slammed into $Kingston_{[E_3]}$ on $monday_{[E_4]}$ with torrential $rains_{[E_5]}$.
S_2: No serious $injuries_{[E_6]}$ were immediately reported in $Kingston_{[E_3]}$.
S_3: For half an $hour_{[E_7]}$, the $hurricane_{[E_1]}$ lashed the $city_{[E_8]}$, tearing $branches_{[E_9]}$ from $trees_{[E_{10}]}$, blowing down $fences_{[E_{11}]}$ and whipping $paper_{[E_{12}]}$ through the $air_{[E_{13}]}$.
S_4: The $National_{[E_{14}]}$ $Weather_{[E_{15}]}$ $Service_{[E_{16}]}$ reported heavy $damage_{[E_{17}]}$ to $Kingston_{[E_3]}$'s $airport_{[E_{18}]}$ and $aircraft_{[E_{19}]}$ parked on its $fields_{[E_{20}]}$.

In each sentence S_i, the selected entities are identified as $[E_i]$. Then, the bipartite entity graph (EG) depicted in Figure 2 can be created taking as input both sentences S_i and entities E_i. In this graph, it can be seen that some nodes representing the sentences are linked to each other by means of an entity. The underlying linguistic intuition here is that entities shared by subsequent sentences contribute to the local coherence of the document. In other words, the main reason for employing an EG for summarization is that it is composed by entity transitions similar to lexical chains [3]. Applying one-mode projections to the EG representing potentially non-adjacent connections between sentences, we obtain the

directed projection graph shown in Fig.3. Contrary to the EG, the projection graph (PG) is directed where the direction of an edges is determined by the order of the sentences in the document. Thus, if two sentences *Si* and *Sj* share an entity, there is an edge between *Si* and *Sj* in that order, but the inverse is not possible. Finally, the $Outdegree(PG)$ centrality measure computes the total number of edges outgoing each node in PG. This coherence score is used to select sentences for a summary in the optimization step, in which the ILP-based model for summarization (Section 3.6) will only select those sentences that maximize both informativeness and the local coherence.

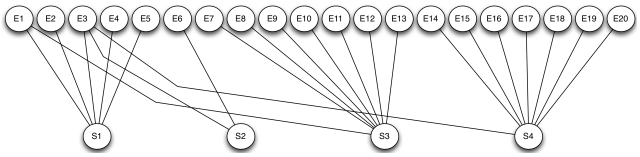

Figure 2: EG representation of the sentences S1-S4 listed above.

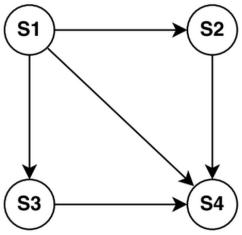

Figure 3: One-mode projection of the bipartite EG

3.6 Summary Generation

For generating a summary, the results of all the previous steps are integrated into an ILP-based optimization model formally described by the objective function (equation 4) and its constraints (equations 5-8). More specifically, the first part of Equation 4 evaluates the score of a concept, while its second part estimate the local coherence of the summaries. In other words, the summary generation task performed in this step is cast as an optimization problem in which the sentences satisfying the previous equations will compose the desired summary. In this model, a sentence in the document to be summarized is coded as a binary value in which 1 indicates that the sentence was selected for composing the summary and 0 that it was not. The proposed ILP-based solution extends the original version introduced in [9] with the *Rank*() term in the objective function. More precisely, the *Rank*() term denotes the coherence score based on the shared entities among sentences captured by the Entity Graph model. Thus, both informativeness and coherence of the text are treated simultaneously by the optimization model denoted by the equations (4-8) bellow:

$$Maximize: \sum_{c_i \in C} w_i \cdot c_i + \sum_{s_j \in S} Rank(s_j) \cdot s_j \quad (4)$$

$$\sum_{s_j \in S} l_j \cdot s_j \leq L \quad (5)$$

$$s_j Occ_{ij} \leq c_i \quad (6)$$

$$\sum_{s_j \in S} s_j Occ_{ij} \geq c_i \quad (7)$$

$$c_j, s_j, Occ_{ij} \in \{0,1\} \forall_{i,j} \quad (8)$$

The variables c_j, s_j and Occ_{ij} are binary values (constraint 8) that indicate a concept c_j, a sentence s_j and the occurrence of a concept c_j in a sentence s_j, respectively. The variable w_i represents the weight of a concept, i.e. the importance of a concept c_j in the set of all the concepts C extracted from the input document. $Rank(s_j)$ is the coherence score of each sentence s_j in the set of sentences S.

The first part $\sum_{c_i \in C} w_i \cdot c_i$ of the objective function defines the importance of the summary, selecting the largest number of important concepts, while the second part $\sum_{s_j \in S} Rank(s_j) \cdot s_j$ is related to coherence of the sentences. The variable l_j is the length of each sentence s_j of the set of sentences S. L is the threshold used to define the maximum length of the generated summary. The inequality 5 ensures that the summary to be generated can not exceed the L value that defines the maximum length of the generated summary. The inequalities 6 and 7 associate the sentences and concepts. This ensures that selecting a sentence leads to selection of all the concepts it contains, and a concept is only selected when it is present in at least one of the selected sentences. Therefore, to ensure the consistency of the ILP model introduced above, the restrictions imposed by equations (4-8) have to be satisfied. In our implementation, this optimization problem is solved by using the GNU Linear Programming Kit (GLPK)[1].

4 EXPERIMENTAL EVALUATION

This section first presents the summarization benchmark datasets and measures used in the experimental evaluation. Then, the results of several experiments aimed at assessing the effectiveness of the proposed summarizer are discussed.

4.1 Datasets and Evaluation Measures

Datasets. The DUC conferences [17] of 2001 and 2002 focused on the generic single-document summarization of news articles written in English and proposed two datasets for single-document summarization. Both datasets are still the most used for evaluating new approaches to single-document summarization. They were created by humans who read the original documents and then produced the *golden summaries*, or reference summaries which are in fact abstractive summaries with approximately 100 words each one. Compared to the summaries produced by a machine, abstractive summaries created by humans usually have high quality. Table 1 summarizes some basic statistics of the DUC datasets.

Evaluation Measures. Evaluating the performance of a summarization system is a difficult task by itself and there exists various evaluation measures that can be used to determine how good a summary is with respect to golden standard summaries. The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [11] consists

[1]https://www.gnu.org/software/glpk/

Table 1: Datasets used in the experiments.

Dataset	N. of Docs	N. of Sent	N. of Words
DUC 2001	309	11.026	269.990
DUC 2002	576	14.370	348.012

in a set of measures available as a software package for automatic evaluation of summaries. Using ROUGE package, it is possible to compare automatically a given generated summary against a set of golden (human-produced) summaries. The ROUGE package provides the ROUGE-N which consists of an n-gram recall between a candidate summary, generated by an automatic text summarization system, and a set of reference summaries called golden standard. In this work, we adopted the most commonly used recall ROUGE-1 (R-1) and ROUGE-2 (R-2) measures. Such measures compute the number of overlapping unigrams and bigrams between the generated summary and the golden standard summaries, respectively. Another advantage of R-1 and R-2 measures is that they have high correlation with human evaluations [11].

4.2 Concept Scoring Evaluation

This experiment evaluates the performance of each concept scoring method (TF, NTF, and TF-ISF). It aims to select the best method to be part of the algorithm that is responsible for weighting concepts. For all of the experiments in this section, the second part of the objective function presented in 4, which denotes the coherence of the sentences, was ignored. The concepts scoring method introduced in Section 3 are evaluated using two distinct weight distribution strategies: *All Concepts (AC)* and *Highest Ranked Concepts (HRC)*. The AC strategy consists of the common approach to assign weights to all concepts in a document, whereas HRC one only assigns weight to concepts belonging to the ratio θ of the top highest ranked concepts. For instance, if $\theta = 1/2$ it will select half of the highest weighted concepts, and $\theta = 1/3$ denotes one third of the highest weighted concepts.

The results yielded in this experiment, in terms of R-1 and R-2, are presented in Table 2. It shows that the HRC strategy which limits the number of extracted concepts from the documents to a given threshold $\theta = 1/3$ yielded best results on both DUC datasets. Overall, regarding the methods of weighing the concepts, the NTF method outperformed the others in most of the cases.

4.3 Evaluation of the Proposed Summarizer

Due to the fact that the proposed summarizer is highly customizable, many components of the system can be tuned to obtain higher performance depending on the dataset, for instance. Thus, the experiment reported in this section has the main goal of selecting the best parameter setting when evaluated on both DUC 2001 and 2002 datasets.

Table 3 shows system settings obtained when one combines the concept scoring method with the EG on the DUC 2001 and 2002 datasets. For all these settings, the threshold $\theta = 1/3$ was adopted since it yielded the best results in the previous experiments. One can notice that, for the both DUC datasets, the best performance in terms of R-1 was achieved by the experimental setting NTF+Entity Graph. Furthermore, the performance difference between this setting was statistically significant, mainly when compared against the experimental setting that not used the Entity Graph for the DUC 2001 dataset. This is an evidence of the positive contribution of the EG in our overall proposed solution for single-document summarization. Although less evident for the DUC 2002 dataset, the contribution of the EG to the overall R-1 and R-2 scores is still positive. On the DUC 2001 dataset, the TF-ISF yielded the highest R-2 score. Finally, comparing the concept scoring methods individually, it can be observed that NTF has outperformed the other two concept scoring methods.

Table 2: Performance of the scoring methods using ILP on DUC 2001 and 2002. The highest R-1 and R-2 scores are in bold.

DUC 2001

Dataset	Documents	R-1	R-2
TF	All occurences	42.91	16.92
NTF	All occurences	43.06	17.11
TF-ISF	All occurences	43.18	17.25
TF	($\theta = 1/3$)	42.91	16.92
NTF	($\theta = 1/3$)	43.15	**17.61**
TF-ISF	($\theta = 1/3$)	**43.44**	17.47

DUC 2002

Dataset	Documents	R-1	R-2
TF	All occurences	45.68	19.39
NTF	All occurences	45.90	19.73
TF-ISF	All occurences	43.13	18.06
TF	($\theta = 1/3$)	45.78	19.54
NTF	($\theta = 1/3$)	**46.11**	**19.99**
TF-ISF	($\theta = 1/3$)	44.53	18.41

Table 3: Performance results combining several features on DUC 2001 and 2002. The highest R-1 and R-2 scores are in bold.

DUC 2001

Experimental Setting	R-1	R-2
TF + Entity Graph	42.89	16.89
NTF + Entity Graph	**45.00**	17.91
TF-ISF + Entity Graph	43.90	**18.27**

DUC 2002

Experimental Setting	R-1	R-2
TF + Entity Graph	46.07	19.87
NTF + Entity Graph	**47.36**	**20.96**
TF-ISF + Entity Graph	45.28	20.17

4.4 Comparative Evaluation

The experiments in this section compare the performance of the proposed solution against the following baseline and other extractive summarizers on both DUC datasets: (i) a baseline defined by selecting the first sentences from the input document up to 100 words; (ii) the best systems participating in the DUC 2001 and 2002

Table 4: Comparative results on DUC 2001 and 2002. The highest R-1 and R-2 scores are in bold.

DUC 2001

Summarizer	R-1	R-2
AutoSummarizer	41.92	16.63
Classifier4J	44.44	19.86
HP-UFPE FS	35.91	11.78
System T	44.53	**20.27**
TextRank	40.66	15.09
Proposed summarizer	**45.00**	17.91

DUC 2002

Summarizer	R-1	R-2
AutoSummarizer	43.79	19.17
Classifier4J	47.09	22.12
HP-UFPE FS	45.70	20.55
System 28	48.07	22.88
TextRank	43.93	18.66
Parveen's summarizer	**48.50**	**23.00**
Proposed summarizer	47.36	20.96

competitions, System T and System 28, respectively; (iii) AutoSummarizer [1], Classifier4J [13], and HP-UFPE FS [8] which had the highest R-1 performance reported in [4]; (iv) TextRank [16]; (v) the extractive summarizer in [19] on DUC 2002.

4.4.1 ROUGE Comparison. Tables 4 show the empirical results of the comparative evaluation in terms of R-1 and R-2 measures. Our summarizer setting which achieved the best performance in the previous experiments (Table 3) were compared to the selected systems.

On the DUC 2001 dataset, the proposed system outperformed the others in terms of R-1 score, whereas System T obtained the best R-2 score. On the DUC 2002, the Parveen's Summarizer achieved the highest scores for both R-1 and R-2 scores, followed by System 28 and the proposed summarizer. These results show that the proposed summarizer presents competitive performance against other state-of-the-art summarizers.

Surprisingly, even after more than a decade since the DUC 2001-2002 competitions, the more recently proposed summarizers evaluated in this paper do not substantially outperform the baseline performance achieved by Systems T and System 28. A possible explanation that is mainly due to the fact that the latter systems were developed specifically for the DUC corpora of their respective competitions. Therefore, more experiments using other datasets, and possibly, on distinct domains should be more clarifying about the actual performance difference among the systems compared here.

4.4.2 Summary Coherence Comparison. In the previous section, we have assessed the summary informativeness using ROUGE measures. However, it is well known that this measure does not correlate well with the coherence rating since other text properties including readability, fluency, and coherence are not taken into account [10]. We first intend to conduct a detailed comparative coherence analysis, but unfortunately the summaries of some summarizers studied in this work are not publicly available. Despite of all that and aiming at showing some evidences that the proposed method can take text coherence into account, we performed an informal human comparison of the coherence level of three summaries produced by our summarizer, Classifier4J, and AutoSummarizer. The three documents used in this comparative assessment were randomly taken from the DUC 2001 corpus. More specifically, we chose one document from each of the following sizes: short (10-30), medium (31-70), and long (>70) sentences. Such a restriction may provide a better understanding of the coherence level of the summaries in function of the input document size. The aforementioned summaries and the comparative discussion are presented next.

Summary 1: Doc. AP891028-0022/DUC 2001 - 18 sentences

Classifier4J
S1: A major earthquake registering 7.2 on the Richter scale shook the Solomon Islands in the South Pacific today, the U.S. Geological Survey says.
S2: The preliminary reading of 7.2 is slightly stronger than the 7.1 magnitude earthquake that hit the San Francisco Bay area Oct. 17.
S3: The earthquake struck the islands at 8:05 a.m. EDT Friday, said USGS spokesman Don Finley.
S4: It was the largest earthquake in the Solomons since a 7.4 quake on Nov. 5, 1978.
S5: There were no immediate reports of injury or damage.
S6: Saturday's earthquake was the strongest in the world in five months, Finley said.

AutoSummarizer
S1: Major earthquakes in the Solomons usually dont cause much damage or many casualties because the area is sparsely populated and not extensively developed.
S2: The surveys earthquake monitors in Golden, Colo., said early seismograph readings placed the epicenter of Saturdays earthquake about 200 miles southeast of Honiaria, which is on Guadacanal Island and is the capital of the Solomons.
S3: An earthquake of 3.5 on the Richter scale can cause slight damage in the local area, 4 moderate damage, 5 considerable damage, 6 severe damage.
S4: A 7 reading is a "major earthquake", capable of widespread heavy damage; 8 is a "great quake", capable of tremendous damage.

Proposed System
S1: A major earthquake registering 7.2 on the Richter scale shook the Solomon Islands in the South Pacific today, the U.S. Geological Survey says.
S2: The preliminary reading of 7.2 is slightly stronger than the 7.1 magnitude earthquake that hit the San Francisco Bay area Oct. 17.
S3: The earthquake struck the islands at 8:05 a.m. EDT Friday, said USGS spokesman Don Finley.
S4: It was the largest earthquake in the Solomons since a 7.4 quake on Nov. 5, 1978.
S5: An earthquake of 3.5 on the Richter scale can cause slight damage in the local area, 4 moderate damage, 5 considerable damage, 6 severe damage.

Summary 2: Doc. FT933-8272/DUC 2001 - 35 sentences

Classifier4J
S1: The epidemic of bovine spongiform encephalopathy or 'mad cow' disease - which has killed more than 100,000 animals in the UK - is causing a new wave of public concern.
S2: New cases are still running at almost 1,000 a week and last month a second dairy farmer died of Creutzfeld-Jacob Disease, a brain disorder similar to BSE.
S3: Richard Lacey, a microbiology professor at Leeds University and the leading critic of government policy on BSE, said the deaths this year of two farmers whose herds had suffered from mad cow disease could not be put down to chance.
S4: I find it unbelievable that the government and their hand-picked advisers can go on telling the public there is no danger.

AutoSummarizer
S1: New cases are still running at almost 1,000 a week and last month a second dairy farmer died of Creutzfeld-Jacob Disease, a brain disorder similar to BSE.
S2: One argument put forward by the health department is that CJD has such a long incubation period - typically 10 to 20 years - that clinical symptoms would not yet have appeared, even if BSE had triggered any cases of CJD. It is most unlikely therefore that there is any direct link between the cases of BSE and the occurrence of disease in the patient.
S3: Most independent experts maintain that no human being - dairy farmer or beef eater - is likely to be exposed to BSE in sufficient quantities to develop brain disease.

Proposed System
S1: The epidemic of bovine spongiform encephalopathy or 'mad cow' disease - which has killed more than 100,000 animals in the UK - is causing a new wave of public concern.
S2: The official view is that the deaths are an unfortunate coincidence, even though it is statistically unlikely that two dairy farmers should contract a disease as rare as CJD.
S3: The source of infection was protein-rich cattle feed contaminated with scrapie, a related brain disease of sheep.
S4: The incubation period is also longer than originally expected.
S5: Veterinary experts say that almost all of the 102,000 confirmed BSE cases so far can be attributed to scrapie-contaminated feed.

Summary 3: Doc. WSJ910107-0139/DUC 2001 - 102 sentences

Classifier4J
S1: Under a microscope, parts of his brain, riddled with little holes, looked like a sponge.
S2: Thus began another chapter in one of medicine's most bizarre mysteries, a tale of sick sheep and mad cows, cannibals and Pennsylvanians, ancient life forms and a cat named Max.
S3: The plot revolves around a family of brain diseases, probably variations of a single disorder, called spongiform encephalopathy.
S4: Spongiform research already has raised questions about a cornerstone of biology and spawned a Nobel Prize.
S5: The uncanny nature of the disorder sometimes grips scientists with a kind of obsessive fascination, notes NIH researcher D. Carleton Gajdusek.

AutoSummarizer
S1: Yet sheep have never been strongly implicated in cases of the human form of the disease, Creutzfeldt-Jakob disease, or CJD. Indeed, CJD is a nefarious trickster; the first reported case of the disease, which entered medical texts in the 1920s, really wasnt CJD after all, scientists now believe.
S2: Based on the clues, Dr. Gajdusek, back at NIH, led studies in the 1960s showing kuru, CJD and scrapie to be essentially the same infectious disease, studies that won the 1976 Nobel Prize for medicine.
S3: Brain tissue of infected animals could transmit the disease when injected into different animals brains, yet microscopes revealed no signs of infectious microbes.
S4: Still, Dr. Brown and other scientists arent much worried about mad cows because their animal studies show spongiform disease is very hard to transmit orally.

Proposed System
S1: The plot revolves around a family of brain diseases, probably variations of a single disorder, called spongiform encephalopathy.
S2: Scientists agree, however, that the disease in animals probably poses little danger to people.
S3: But the infectious agent continued to baffle scientists.
S4: It seemed like "biological spontaneous combustion," says NIH researcher Paul Brown.
S5: Now more than 10 British cats have died of it, suggesting brains from infected sheep or cows got into cat food.
S6: During the past few months, the NIH team, working with Dr. Mitrova, has found a thread that for the first time may link many such CJD cases.

Discussion. In Summary 1, the summaries produced by Classifier4J and the proposed system are highly coherent. Indeed, both summarizers selected the same initial sentences ($S1$-$S4$), differing only in the last two. A possible reason resides in the fact that the shorter the input document is, the more probable is to obtain a reasonable coherent summary. On the other hand, AutoSummarizer seems to prefer longer sentences to be included in the summaries it generates. As a result, its summary clearly breaks the discourse flow from the very beginning (sentence $S1$ and $S2$).

In Summary 2, the proposed system agrees with Classifier4J in only one sentence ($S1$). Despite of that, both summarizers produced comparable coherent summaries. Again, the longer sentences selected by AutoSummarizer makes the reading confusing starting from the very first sentence.

Finally, concerning longer input documents (Summary 3), one can notice that the summaries produced by Classifier4J and AutoSummarizer have disconnected sentences at the beginning, while the rest of the sentences shows superior coherence among them. On the contrary, the last summary generated by the proposed system clearly has superior coherence among the sentences, i.e., the reading is easy to follow.

5 CONCLUSION AND FUTURE WORK

This paper presented an extractive single-document summarizer that combines unsupervised methods for concept extraction, concept scoring, and coherence scoring. Moreover, the proposed solution integrates all of the above methods in a single objective function that can not only maximize the relevance of the summaries, but also improve their coherence. Our solution was evaluated and compared with state-of-the-art summarizers on two benchmark datasets. The achieved results in terms of R-1 and R-2 measures were encouraging and competitive showing that our solution is effective in producing good summaries. A preliminary human-based evaluation also shown that the summaries generated by our summarizer was usually more coherent than summaries produced by two other summarizers.

Our extractive summarizer is a generic one, i.e., its summarization process (mainly concept extraction and scoring) is the same for any type of input document regardless of its domain. However, as shown in [7], different scoring techniques are usually better suited to a specific domain. Based on this evidence, we intend to conduct a deeper analysis of the overall impact of both concept and coherence scoring in different domains including articles in Computer Science, and Biomedicine. We are aware of the space left by this work regarding a sound evaluation methodology of the summary coherence aspects. For that, a study of both summary coherence and readability ratings, as done in [10], is been conducted. In this scenario, the readability rating will provide an estimate of the interpretation effort needed by the reader. Furthermore, inspired by [12], we will examine how discourse relations can be integrated into our summarizer. Finally, further work to extend the proposed summarizer to the multi-document scenario will be undertaken. For that, other important aspects besides the overall coherence of the summary should be addressed, including redundancy control, and sentence ordering.

REFERENCES

[1] 2016. AutoSummarizer, http://autosummarizer.com. (2016).

[2] Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. 2016. Multi-document abstractive summarization using ILP based multi-sentence compression. *CoRR* abs/1609.07034 (2016).

[3] Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. In *Proceedings of the ACL/EACL 1997 Workshop on Intelligent Scalable Text Summarization*. 10–17.

[4] Jamilson Batista, Rodolfo Ferreira, Hilário Tomaz, Rafael Ferreira, Rafael Dueire Lins, Steven Simske, Gabriel Silva, and Marcelo Riss. 2015. A Quantitative and Qualitative Assessment of Automatic Text Summarization Systems. In *Proceedings of the 2015 ACM Symposium on Document Engineering (DocEng '15)*. ACM, New York, NY, USA, 65–68.

[5] Catherine Blake. 2006. A Comparison of Document, Sentence, and Term Event Spaces. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (ACL-44)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 601–608.

[6] Florian Boudin, Hugo Mougard, and Benot Favre. 2015. Concept-based Summarization using Integer Linear Programming: From Concept Pruning to Multiple Optimal Solutions. In *EMNLP*, Llus Mrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton (Eds.). The Association for Computational Linguistics, 1914–1918.

[7] Rafael Ferreira, Luciano De Souza Cabral, Rafael Dueire Lins, Gabriel Pereira E Silva, Fred Freitas, George D C Cavalcanti, Rinaldo Lima, Steven J. Simske, and Luciano Favaro. 2013. Assessing sentence scoring techniques for extractive text summarization. *Expert Systems with Applications* 40, 14 (2013), 5755–5764.

[8] Rafael Ferreira, Frederico Freitas, Luciano de Souza Cabral, Rafael Dueire Lins, Rinaldo Lima, Gabriel Franca, Steven J. Simske, and Luciano Favaro. 2014. A Context Based Text Summarization System. *2014 11th IAPR International Workshop on Document Analysis Systems* (2014), 66–70.

[9] Dan Gillick and Benoit Favre. 2009. A Scalable Global Model for Summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing (ILP '09)*. Association for Computational Linguistics, 10–18.

[10] Camille Guinaudeau and Michael Strube. 2013. Graph-based Local Coherence Modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, 93–103. http://www.aclweb.org/anthology/P13-1010

[11] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, Stan Szpakowicz Marie-Francine Moens (Ed.). Association for Computational Linguistics, Barcelona, Spain, 74–81.

[12] Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (Eds.). 2011. *Automatically evaluating text coherence using discourse relations. The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA.*

[13] N Lothian. 2003. Classifier4J, http://classifier4j.sourceforge.net. (2003). http://classifier4j.sourceforge.net/

[14] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL (System Demonstrations)*. The Association for Computer Linguistics, 55–60.

[15] Ryan T. McDonald. 2007. A Study of Global Inference Algorithms in Multi-document Summarization. In *ECIR* (2009-07-02) *(Lecture Notes in Computer Science)*, Giambattista Amati, Claudio Carpineto, and Giovanni Romano (Eds.), Vol. 4425. Springer, 557–564.

[16] R. Mihalcea and P. Tarau. 2004. TextRank: Bringing Order into Texts. In *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*.

[17] Paul Over, Hoa Dang, and Donna Harman. 2007. DUC in Context. *Inf. Process. Manage.* 43, 6 (Nov. 2007), 1506–1520.

[18] Daraksha Parveen, Hans-Martin Ramsl, and Michael Strube. 2015. Topical Coherence for Graph-based Extractive Summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 1949–1954.

[19] Daraksha Parveen and Michael Strube. 2015. Integrating Importance, Non-redundancy and Coherence in Graph-based Extractive Summarization. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*. AAAI Press, 1298–1304. http://dl.acm.org/citation.cfm?id=2832415.2832430

[20] Casper Petersen, Christina Lioma, Jakob Grue Simonsen, and Birger Larsen. 2015. Entropy and Graph Based Modelling of Document Coherence Using Discourse Entities: An Application to IR. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval (ICTIR '15)*. ACM, New York, NY, USA, 191–200. https://doi.org/10.1145/2808194.2809458

[21] Gerard Salton and Christopher Buckley. 1988. Term-weighting Approaches in Automatic Text Retrieval. *Inf. Process. Manage.* 24, 5 (Aug. 1988), 513–523.

[22] Natalie Schluter and Anders Søgaard. 2015. Unsupervised extractive summarization via coverage maximization with syntactic and semantic concepts. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Vol. 2. Association for Computational Linguistics, 840–844.

[23] Karin Sim Smith, Wilker Aziz, and Lucia Specia. 2016. Cohere: A Toolkit for Local Coherence. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (23-28), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Paris, France.