A Language Independent Algorithm for Single and Multiple Document Summarization

Rada Mihalcea and **Paul Tarau**
Department of Computer Science and Engineering
University of North Texas
{rada,tarau}@cs.unt.edu

Abstract

This paper describes a method for *language independent* extractive summarization that relies on iterative graph-based ranking algorithms. Through evaluations performed on a single-document summarization task for English and Portuguese, we show that the method performs equally well regardless of the language. Moreover, we show how a meta-summarizer relying on a layered application of techniques for single-document summarization can be turned into an effective method for multi-document summarization.

1 Introduction

Algorithms for extractive summarization are typically based on techniques for sentence extraction, and attempt to identify the set of sentences that are most important for the overall understanding of a given document. Some of the most successful approaches consist of supervised algorithms that attempt to learn what makes a good summary by training on collections of summaries built for a relatively large number of training documents, e.g. (Hirao et al., 2002), (Teufel and Moens, 1997). However, the price paid for the high performance of such supervised algorithms is their inability to easily adapt to new languages or domains, as new training data are required for each new data type. In this paper, we show that a method for extractive summarization relying on iterative graph-based algorithms, as previously proposed in (Mihalcea and Tarau, 2004) can be applied to the summarization of documents in different languages without any requirements for additional data. Additionally, we also show that a layered application of this single-document summarization method can result into an efficient multi-document summarization tool.

Earlier experiments with graph-based ranking algorithms for text summarization, as previously reported in (Mihalcea and Tarau, 2004) and (Erkan and Radev, 2004), were either limited to single-document English summarization, or they were applied to English multi-document summarization, but in conjunction with other extractive summarization techniques that did not allow for a clear evaluation of the impact of the graph algorithms alone. In this paper, we show that a method exclusively based on graph-based algorithms can be successfully applied to the summarization of single and multiple documents in any language, and show that the results are competitive with those of state-of-the-art summarization systems.

The paper is organized as follows. Section 2 briefly overviews two iterative graph-based ranking algorithms, and shows how these algorithms can be applied to single and multiple document summarization. Section 3 describes the data sets used in the summarization experiments and the evaluation methodology. Experimental results are presented in Section 4, followed by discussions, pointers to related work, and conclusions.

2 Iterative Graph-based Algorithms for Extractive Summarization

In this section, we shortly describe two graph-based ranking algorithms and their application to the task of extractive summarization. Ranking algorithms, such as Kleinberg's HITS algorithm (Kleinberg, 1999) or Google's PageRank (Brin and Page, 1998), have been traditionally and successfully used in Web-link analysis (Brin and Page, 1998), social networks, and more recently in text processing applications (Mihalcea and Tarau, 2004), (Mihalcea et al., 2004), (Erkan and Radev, 2004). In short, a graph-based ranking algorithm is a way of deciding on the importance of a vertex within a graph, by taking into account global information recursively computed from the entire graph, rather than relying

only on local vertex-specific information. The basic idea implemented by the ranking model is that of "voting" or "recommendation". When one vertex links to another one, it is basically casting a vote for that other vertex. The higher the number of votes that are cast for a vertex, the higher the importance of the vertex.

Let $G = (V, E)$ be a directed graph with the set of vertices V and set of edges E, where E is a subset of $V \times V$. For a given vertex V_i, let $In(V_i)$ be the set of vertices that point to it (predecessors), and let $Out(V_i)$ be the set of vertices that vertex V_i points to (successors).

PageRank. $PageRank$ (Brin and Page, 1998) is perhaps one of the most popular ranking algorithms, and was designed as a method for Web link analysis. Unlike other graph ranking algorithms, $PageRank$ integrates the impact of both incoming and outgoing links into one single model, and therefore it produces only one set of scores:

$$PR(V_i) = (1 - d) + d * \sum_{V_j \in In(V_i)} \frac{PR(V_j)}{|Out(V_j)|} \quad (1)$$

where d is a parameter set between 0 and 1.

HITS. $HITS$ (Hyperlinked Induced Topic Search) (Kleinberg, 1999) is an iterative algorithm that was designed for ranking Web pages according to their degree of "authority". The $HITS$ algorithm makes a distinction between "authorities" (pages with a large number of incoming links) and "hubs" (pages with a large number of outgoing links). For each vertex, $HITS$ produces two sets of scores – an "authority" score, and a "hub" score:

$$HITS_A(V_i) = \sum_{V_j \in In(V_i)} HITS_H(V_j) \quad (2)$$

$$HITS_H(V_i) = \sum_{V_j \in Out(V_i)} HITS_A(V_j) \quad (3)$$

For each of these algorithms, starting from arbitrary values assigned to each node in the graph, the computation iterates until convergence below a given threshold is achieved. After running the algorithm, a score is associated with each vertex, which represents the "importance" or "power" of that vertex within the graph.

In the context of Web surfing or citation analysis, it is unusual for a vertex to include multiple or partial links to another vertex, and hence the original definition for graph-based ranking algorithms is assuming unweighted graphs. However, when the graphs are built starting with natural language texts, they may include multiple or partial links between the units (vertices) that are extracted from text. It may be therefore useful to integrate into the model the "strength" of the connection between two vertices V_i and V_j as a weight w_{ij} added to the corresponding edge that connects the two vertices. The ranking algorithms are thus adapted to include edge weights, e.g. for $PageRank$ the score is determined using the following formula (a similar change can be applied to the $HITS$ algorithm):

$$PR^W(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} w_{ji} \frac{PR^W(V_j)}{\sum_{V_k \in Out(V_j)} w_{kj}} \quad (4)$$

[1] Watching the new movie, "Imagine: John Lennon," was very painful for the late Beatle's wife, Yoko Ono.
[2] "The only reason why I did watch it to the end is because I'm responsible for it, even though somebody else made it," she said.
[3] Cassettes, film footage and other elements of the acclaimed movie were collected by Ono.
[4] She also took cassettes of interviews by Lennon, which were edited in such a way that he narrates the picture.
[5] Andrew Solt ("This Is Elvis") directed, Solt and David L. Wolper produced and Solt and Sam Egan wrote it.
[6] "I think this is really the definitive documentary of John Lennon's life," Ono said in an interview.

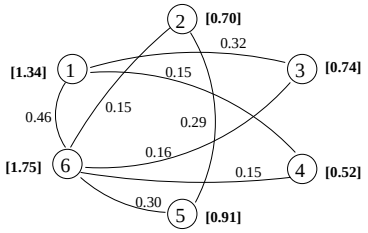

Figure 1: Graph of sentence similarities built on a sample text. Scores reflecting sentence importance are shown in brackets next to each sentence.

While the final vertex scores (and therefore rankings) for weighted graphs differ significantly as compared to their unweighted alternatives, the number of iterations to convergence and the shape of the convergence curves is almost identical for weighted and unweighted graphs.

2.1 Single Document Summarization

For the task of single-document extractive summarization, the goal is to rank the sentences in a given text with respect to their importance for the overall understanding of the text. A graph is therefore constructed by adding a vertex for each sentence in the text, and edges between vertices are established using sentence inter-connections. These connections

are defined using a similarity relation, where "similarity" is measured as a function of content overlap. Such a relation between two sentences can be seen as a process of "recommendation": a sentence that addresses certain concepts in a text gives the reader a "recommendation" to refer to other sentences in the text that address the same concepts, and therefore a link can be drawn between any two such sentences that share common content.

The overlap of two sentences can be determined simply as the number of common tokens between the lexical representations of two sentences, or it can be run through syntactic filters, which only count words of a certain syntactic category. Moreover, to avoid promoting long sentences, we use a normalization factor, and divide the content overlap of two sentences with the length of each sentence.

The resulting graph is highly connected, with a weight associated with each edge, indicating the strength of the connections between various sentence pairs in the text. The graph can be represented as: (a) simple *undirected* graph; (b) directed weighted graph with the orientation of edges set from a sentence to sentences that follow in the text (*directed forward*); or (c) directed weighted graph with the orientation of edges set from a sentence to previous sentences in the text (*directed backward*).

After the ranking algorithm is run on the graph, sentences are sorted in reversed order of their score, and the top ranked sentences are selected for inclusion in the extractive summary. Figure 1 shows an example of a weighted graph built for a sample text of six sentences.

2.2 Multiple Document Summarization

Multi-document summaries are built using a "meta" summarization procedure. First, for each document in a given cluster of documents, a single document summary is generated using one of the graph-based ranking algorithms. Next, a "summary of summaries" is produced using the same or a different ranking algorithm. Figure 2 illustrates the meta-summarization process used to generate a multi-document summary starting with a cluster of N documents.

Unlike single documents – where sentences with highly similar content are very rarely if at all encountered – it is often the case that clusters of multiple documents, all addressing the same or related topics, would contain very similar or even identical sentences. To avoid such pairs of sentences, which may decrease the readability and the amount of information conveyed by a summary, we introduce a maximum threshold on the sentence similarity measure. Consequently, in the graph construction stage, no link (edge) is added between sentences (vertices) whose similarity exceeds this threshold. In the experiments reported in this paper, this similarity threshold was empirically set to 0.5.

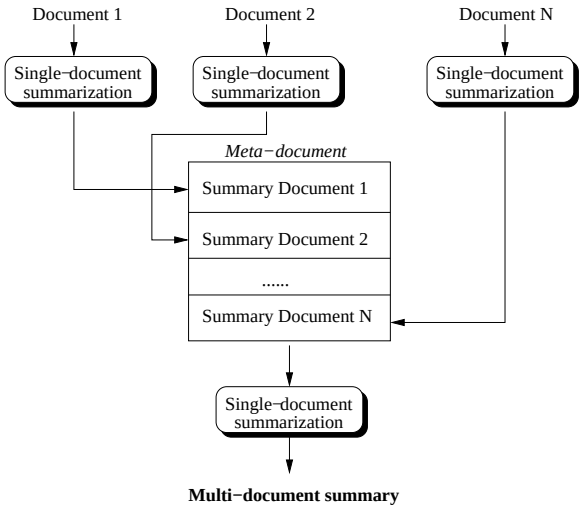

Figure 2: Generation of a multi-document summary using meta-summarization.

3 Materials and Evaluation Methodology

Single and multiple English document summarization experiments are run using the summarization test collection provided in the framework of the Document Understanding Conference (DUC). In particular, we use the data set of 567 news articles made available during the DUC 2002 evaluations (DUC, 2002), and the corresponding 100-word summaries generated for each of these documents (single-document summarization), or the 100-word summaries generated for each of the 59 document clusters formed on the same data set (multi-document summarization). These are the summarization tasks undertaken by other systems participating in the DUC 2002 document summarization evaluations.

To test the language independence aspect of the algorithm, in addition to the English test collection, we also use a Brazilian Portuguese data set consisting of 100 news articles and their corresponding manually produced summaries. We use the TeMário test collection (Pardo and Rino, 2003), containing newspaper articles from online Brazilian newswire: 40 documents from *Jornal de Brasil* and 60 documents from *Folha de São Paulo*. The documents were selected to cover a variety of domains (e.g. world, politics, foreign affairs, editorials), and manual summaries were produced by an expert in Brazilian Portuguese. Unlike the summaries produced for the English DUC documents – which had a length requirement of approximately 100 words, the length of the summaries in the TeMário data

set is constrained relative to the length of the corresponding documents, i.e. a summary has to account for about 25-30% of the original document. Consequently, the automatic summaries generated for the documents in this collection are not restricted to 100 words, as in the English experiments, but are required to have a length comparable to the corresponding manual summaries, to ensure a fair evaluation.

For evaluation, we are using the ROUGE evaluation toolkit[1], which is a method based on Ngram statistics, found to be highly correlated with human evaluations (Lin and Hovy, 2003a). The evaluation is done using the Ngram(1,1) setting of ROUGE, which was found to have the highest correlation with human judgments, at a confidence level of 95%.

4 Experimental Results

The extractive summarization algorithm is evaluated in the context of: (1) A single-document summarization task, where a summary is generated for each of the 567 English news articles provided during the Document Understanding Evaluations 2002 (DUC, 2002), and for each of the 100 Portuguese documents in the TeMário data set; and (2) A multi-document summarization task, where a summary is generated for each of the 59 document clusters in the DUC 2002 data. Since document clusters and multi-document summaries are not available for the Portuguese documents, a multi-document summarization evaluation could not be conducted on this data set. Note however that the multi-document summarization tool is based on the single-document summarization method (see Figure 2), and thus high performance in single-document summarization is expected to result into a similar level of performance in multi-document summarization.

4.1 Single Document Summarization for English

For single-document summarization, we evaluate the extractive summaries produced using each of the two graph-based ranking algorithms described in Section 2 ($HITS$ and $PageRank$). Table 1 shows the results obtained for the 100-words automatically generated summaries for the English DUC 2002 data set. The table shows results using the two graph algorithms described in Section 2 when using graphs that are: (a) undirected, (b) directed forward, or (c) directed backward[2].

[1] ROUGE is available at http://www.isi.edu/~cyl/ROUGE/.

[2] Note that the first two rows in the table are in fact redundant, since the "hub" variation of the HITS algorithm can be derived from its "authority" counterpart by reversing the edge orientation in the graphs.

Algorithm	Graph		
	Undir.	Forward	Backward
$HITS_A^W$	49.12	45.84	**50.23**
$HITS_H^W$	49.12	**50.23**	45.84
$PageRank^W$	49.04	42.02	**50.08**

Table 1: Results for English single-document summarization.

For a comparative evaluation, Table 2 shows the results obtained on this data set by the top 5 (out of 15) performing systems participating in the single document summarization task at DUC 2002. It also lists the baseline performance, computed for 100-word summaries generated by taking the first sentences in each article.

Top 5 systems (DUC, 2002)					
S27	S31	S28	S21	S29	Baseline
50.11	49.14	48.90	48.69	46.81	47.99

Table 2: Results for top 5 DUC 2002 single document summarization systems, and baseline.

4.2 Single Document Summarization for Portuguese

The single-document summarization tool was also evaluated on the TeMário collection of Portuguese newspaper articles. We used the same graph settings as in the English experiments: graph-based ranking algorithms consisting of either $HITS$ or $PageRank$, relying on graphs that are undirected, directed forward, or directed backward. As mentioned in Section 3, the length of each automatically generated summary was constrained to match the length of the corresponding manual summary, for a fair comparison. Table 3 shows the results obtained on this data set, evaluated using the ROUGE evaluation toolkit. A baseline was also computed, using the first sentences in each document, and evaluated at 0.4963.

Algorithm	Graph		
	Undir.	Forward	Backward
$HITS_A^W$	48.14	48.34	50.02
$HITS_H^W$	48.14	50.02	48.34
$PageRank^W$	49.39	45.74	**51.21**

Table 3: Results for Portuguese single-document summarization.

4.3 Multiple Document Summarization

We evaluate multi-document summaries generated using combinations of the graph-based ranking algorithms that were found to work best in the single document summarization experiments – $PageRank^W$ and $HITS_A^W$, on undirected or di-

rected backward graphs. Although the single document summaries used in the "meta" summarization process may conceivably be of any size, in this evaluation their length is limited to 100 words.

As mentioned earlier, different graph algorithms can be used for producing the single document summary and the "meta" summary; Table 4 lists the results for multi-document summarization experiments using various combinations of graph algorithms. For comparison, Table 5 lists the results obtained by the top 5 (out of 9) performing systems in the multi-document summarization task at DUC 2002, and a baseline generated by taking the first sentence in each article.

Since no multi-document clusters and associated summaries were available for the other language considered in our experiments, the multi-document summarization experiments were conducted only on the English data set. However, since the multi-doc summarization technique consists of a layered application of single-document summarization, we believe that the performance achieved in single-document summarization for Portuguese would eventually result into similar performance figures when applied to the summarization of clusters of documents.

Top 5 systems (DUC, 2002)					
S26	S19	S29	S25	S20	Baseline
35.78	34.47	32.64	30.56	30.47	29.32

Table 5: Results for top 5 DUC 2002 multi-document summarization systems, and baseline.

4.4 Discussion

The graph-based extractive summarization algorithm succeeds in identifying the most important sentences in a text (or collection of texts) based on information exclusively drawn from the text itself. Unlike other supervised systems, which attempt to learn what makes a good summary by training on collections of summaries built for other articles, the graph-based method is fully unsupervised, and relies only on the given texts to derive an extractive summary.

For single document summarization, the $HITS_A^W$ and $PageRank^W$ algorithms, run on a graph structure encoding a backward direction across sentence relations, provide the best performance. These results are consistent across languages – with similar performance figures observed on both the English DUC data set and on the Portuguese TeMário data set. The setting that is always exceeding the baseline by a large margin is $PageRank^W$ on a directed backward graph, with clear improvements over the simple (but powerful) first-sentence selection baseline. Moreover, comparative evaluations performed with respect to other systems participating in the DUC 2002 evaluations revealed the fact that the performance of the graph-based extractive summarization method is competitive with state-of-the-art summarization systems.

Interestingly, the "directed forward" setting is consistently performing worse than the baseline, which can be explained by the fact that both data sets consist of newspaper articles, which tend to concentrate the most important facts toward the beginning of the document, and therefore disfavor a forward direction set across sentence relations.

For multiple document summarization, the best "meta" summarizer is the $PageRank^W$ algorithm applied on undirected graphs, in combination with a single summarization system using the $HITS_A^W$ ranking algorithm, for a performance similar to the one of the best system in the DUC 2002 multi-document summarization task.

The results obtained during all these experiments prove that graph-based ranking algorithms, previously found successful in Web link analysis and social networks, can be turned into a state-of-the-art tool for extractive summarization when applied to graphs extracted from texts. Moreover, the method was also shown to be language independent, leading to similar results when applied to the summarization of documents in different languages.

The better results obtained by algorithms like $HITS_A^W$ and PageRank on graphs containing only backward edges are likely to come from the fact that recommendations flowing toward the beginning of the text take advantage of the bias giving higher summarizing value of sentences occurring at the beginning of the document.

Another important aspect of the method is that it gives a ranking over all sentences in a text (or a collection of texts) – which means that it can be easily adapted to extracting very short summaries, or longer more explicative summaries.

4.5 Related Work

Extractive summarization is considered an important first step for more sophisticated automatic text summarization. As a consequence, there is a large body of work on algorithms for extractive summarization undertaken as part of the DUC evaluation exercises (http://www-nlpir.nist.gov/projects/duc/). Previous approaches include supervised learning (Hirao et al., 2002), (Teufel and Moens, 1997), vectorial similarity computed between an initial abstract and sentences in the given document, intra-document similarities (Salton et al., 1997), or graph algorithms (Mihalcea and Tarau, 2004), (Erkan and Radev, 2004), (Wolf and Gibson, 2004). It is also

Single document summarization algo.	"Meta" summarization algorithm			
	$PageRank^W$-U	$PageRank^W$-DB	$HITS_A^W$-U	$HITS_A^W$-DB
$PageRank^W$-U	**35.52**	34.99	34.56	34.65
$PageRank^W$-DB	35.02	34.48	35.19	34.39
$HITS_A^W$-U	33.68	32.59	32.12	34.23
$HITS_A^W$-DB	**35.72**	35.20	34.62	34.73

Table 4: Results for multi-document summarization (U = Undirected; DB = Directed Backward)

notable the study reported in (Lin and Hovy, 2003b) discussing the usefulness and limitations of automatic sentence extraction for text summarization, which emphasizes the need of accurate tools for sentence extraction as an integral part of automatic summarization systems.

5 Conclusions

Intuitively, iterative graph-based ranking algorithms work well on the task of extractive summarization because they do not only rely on the local context of a text unit (vertex), but they rather take into account information recursively drawn from the entire text (graph). Through the graphs it builds on texts, a graph-based ranking algorithm identifies connections between various entities in a text, and implements the concept of *recommendation*. A text unit recommends other related text units, and the strength of the recommendation is recursively computed based on the importance of the units making the recommendation. In the process of identifying important sentences in a text, a sentence recommends another sentence that addresses similar concepts as being useful for the overall understanding of the text. Sentences that are highly recommended by other sentences are likely to be more informative, and will be therefore given a higher score.

In this paper, we showed that a previously proposed method for graph-based extractive summarization can be successfully applied to the summarization of documents in different languages, without any requirements for additional knowledge or corpora. Moreover, we showed how a meta-summarizer relying on a layered application of techniques for single-document summarization can be turned into an effective method for multi-document summarization. Experiments performed on standard data sets have shown that the results obtained with this method are comparable with those of state-of-the-art systems for automatic summarization, while at the same time providing the benefits of a robust language independent algorithm.

Acknowledgments

We are grateful to Lucia Helena Machado Rino for making available the TeMário summarization test collection and for her help with this data set.

References

S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7).

DUC. 2002. Document understanding conference 2002. http://www-nlpir.nist.gov/projects/duc/.

G. Erkan and D. Radev. 2004. Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, July.

T. Hirao, Y. Sasaki, H. Isozaki, and E. Maeda. 2002. Ntt's text summarization system for duc-2002. In *Proceedings of the Document Understanding Conference 2002*.

J.M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

C.Y. Lin and E.H. Hovy. 2003a. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of Human Language Technology Conference (HLT-NAACL 2003)*, Edmonton, Canada, May.

C.Y. Lin and E.H. Hovy. 2003b. The potential and limitations of sentence extraction for summarization. In *Proceedings of the HLT/NAACL Workshop on Automatic Summarization*, Edmonton, Canada, May.

R. Mihalcea and P. Tarau. 2004. TextRank – bringing order into texts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*, Barcelona, Spain.

R. Mihalcea, P. Tarau, and E. Figa. 2004. PageRank on semantic networks, with application to word sense disambiguation. In *Proceedings of the 20st International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland.

T.A.S. Pardo and L.H.M. Rino. 2003. TeMario: a corpus for automatic text summarization. Technical report, NILC-TR-03-09.

G. Salton, A. Singhal, M. Mitra, and C. Buckley. 1997. Automatic text structuring and summarization. *Information Processing and Management*, 2(32).

S. Teufel and M. Moens. 1997. Sentence extraction as a classification task. In *ACL/EACL workshop on "Intelligent and scalable Text summarization"*, pages 58–65, Madrid, Spain.

F. Wolf and E. Gibson. 2004. Paragraph-, word-, and coherence-based approaches to sentence ranking: A comparison of algorithm and human performance. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*, Barcelona, Spain, July.