2014 11th IAPR International Workshop on Document Analysis Systems

A Context Based Text Summarization System

Rafael Ferreira*†, Frederico Freitas*, Luciano de Souza Cabral*, Rafael Dueire Lins*, Rinaldo Lima*,
Gabriel França*, Steven J. Simske‡, and Luciano Favaro§
*Informatics Center, Federal University of Pernambuco, Recife, Pernambuco, Brazil
†Department of statistics and informatics, Federal Rural University of Pernambuco, Recife, Pernambuco, Brazil
‡Hewlett-Packard Labs., Fort Collins, CO 80528, USA
§Hewlett-Packard Brazil, Barueri, São Paulo, Brazil

***Abstract*—Text summarization is the process of creating a shorter version of one or more text documents. Automatic text summarization has become an important way of finding relevant information in large text libraries or in the Internet. Extractive text summarization techniques select entire sentences from documents according to some criteria to form a summary. Sentence scoring is the technique most used for extractive text summarization, today. Depending on the context, however, some techniques may yield better results than some others. This paper advocates the thesis that the quality of the summary obtained with combinations of sentence scoring methods depend on text subject. Such hypothesis is evaluated using three different contexts: news, blogs and articles. The results obtained show the validity of the hypothesis formulated and point at which techniques are more effective in each of those contexts studied.**

I. Introduction

The massive quantity of data available today on the Internet has reached unforeseen volumes; thus, it is humanly unfeasible to efficiently sieve useful information from it. The demand for automatic tools that are able to "understand", index, classify and present information in a clear and concise way of text documents has grown drastically in recent years. One solution to this problem is using automatic text summarization (TS) techniques.

Text Summarization focuses on getting the "meaning" of documents. Essentially, TS techniques are classified as *Extractive* and *Abstractive* [1]. Extractive summaries produce a set of the most significant sentences from a document, exactly as they appear. Abstractive summaries attempt to improve the coherence among sentences, and may even produce new ones.

Extractive methods are usually performed in three steps [2]: (i) Create an intermediate representation of the original text; (ii) Sentence scoring; and (iii) Selecting a summary consisting of several sentences.

The first of the steps above creates a representation of the document. Usually, it divides the text into paragraphs, sentences, and tokens. Sometimes some preprocessing, such as stop words removal, is also performed. The second step tries to determine which sentences are important to the document or to which extent it combines information about different topics, by sentence scoring. The score measures how relevant a sentence is to the "understanding" of the text as a whole. The last step combines the scores provided by the previous steps and generates a summary.

Depending on the context, however, some techniques may yield better results than some others [3]. This paper advocates the thesis that the quality of the summary obtained with combinations of sentence scoring methods depend on text subject. Such hypothesis is evaluated using three different contexts: news, blogs and articles.

This paper proposes a new summarization system that easily combines different sentence scoring methods in order to obtain the best summaries depending on the context. The fifteen sentence scoring methods most widely used and referenced in the technical literature in the last 10 years [3] are applied to single document summarization.

Three different datasets (with news, blogs and articles) are used to evaluate the validity of the proposed thesis. Both quantitative and qualitative measures are used to evaluate which combination of the sentence scoring methods yield better results for each context. Combining 3 to 5 specific sentences scoring methods in a certain context provides much better quality results. The choice of those methods depend on context of the document.

II. Summarization Systems Combining Sentence Scoring Methods

The main idea of this work is provide an assessment of sentence scoring methods and a user friendly interface to combine them. This section is divided into: (i) sentence scoring methods description and how we implement it; (ii) the mechanism that provides an easy combination of these methods; (iii) the system flow of activities.

A. Sentence Scoring Methods

Fifteen of the most popular sentence scoring methods are used here. A brief overview of each of them and their implementation details are presented here [1]. In general, sentence scoring methods are classified according to three categories: Word-based Scoring, Sentence-based Scoring and Graph-based Scoring. It is important to notice that all services provide as an output a score between 0 and 1 for each sentence. Some implementations employ a score normalization step.

1) Word-based Scoring: The first methods used for sentence scoring were based on word scoring. In such approaches, each word receives a score and the weight of each sentence is the sum of all scores of its constituent words. The most important word-based scoring methods are listed below.

[1] More details about them in [3]

978-1-4799-3243-6/14 $31.00 © 2014 IEEE
DOI 10.1109/DAS.2014.19

- **Word Frequency:** As the name of the method suggests, the more frequently a words occurs in the text, the higher its score;
- **TF/IDF:** It uses TF/IDF formula [4] to score sentences;
- **Word Co-occurrence:** measures the probability of two terms in a text to appear alongside each other in a certain order;
- **Lexical Similarity:** It is based on the assumption that the important sentences are identified by strong chains;
- **Upper Case:** This method assigns higher scores to words that contain one or more upper case letters;
- **Proper Noun:** This method hypothesizes that sentences that contain a higher number of proper nouns are possibly more important than others.

2) Sentence-based Scoring: This approach analyzes the features of the sentence itself, such as the presence of cue expressions. It was used for the first time in 1968 [5]. The most important methods that follow this idea are described below.

- **Cue-Phrases:** In general, the sentences started by "in summary", "in conclusion", "our investigation", "the paper describes" and emphasizes such as "the best", "the most important", "according to the study", "significantly", "important", "in particular", "hardly", "impossible" as well as domain-specific bonus phrases terms can be good indicators of significant content of a text document;
- **Sentence Position:** The position of the sentence in general influences on its importance. For example, the most important sentences tend to come at the beginning of a document;
- **Sentence Resemblance to the Title:** Sentence resemblance to the title is the vocabulary overlap between this sentence and the document title;
- **Sentence Centrality:** Sentence centrality is the vocabulary overlap between a sentence and other sentences in the document;
- **Sentence Length:** This feature is employed to penalize sentences that are either too short or long;
- **Sentence Inclusion of Numerical Data:** Usually the sentence that contains numerical data is an important one and it is very likely to be included in the document summary.

3) Graph-based Scoring: In graph-based methods the score is generated by the relationship among sentences. When a sentence refers to another it generates a link with an associated weight between them. The weights are used to generate the scores of a sentence.

- **Text Rank:** It extracts the important keywords from a text document and also determines the weight of the "importance" of words within the entire document by using a graph-based model;
- **Bushy Path of the Node:** The bushy path of a node (sentence) on a map is defined as the number of links connecting it to other nodes (sentences) on the map;
- **Aggregate Similarity:** Instead of counting the number of links connecting a node (sentence) to other nodes (Bushy Path), aggregate similarity sums the weights (similarities) of the links.

B. Combining Sentence Scoring Methods

Two new ways of combining the sentence scoring methods are proposed: (i) **By Ranking:** Every service selects the main sentences and the user combines it somehow; and **By Punctuation:** The service scores each sentence and returns one sentence with updated scores.

To improve reusability and to ease service instantiation, the *Template Method* design pattern [6] was implemented. All the methods implemented in this module extend the sentence scoring abstract class. There are four methods in this class: (i) **sentenceScoringRanking:** An abstract method to implement the sentence scoring method. The output is a string list with sentences suggested to be included in the summary; (ii) **sentenceScoringPunctuation:** An abstract method to implement the sentence scoring method. The output is a list of sentences with the method scoring set; (iii) **templateSentenceScoringRanking:** The concrete methods which perform some pre and post processing before calling the *sentenceScoringRanking* abstract method; (iv)**templateSentenceScoringPunctuation:** The concrete methods which perform some pre and post processing before calling the *sentenceScoringPuntiation* abstract method.

The first two methods encapsulate the main part of the service. Thus, all services must implement them. The *templateSentenceScoringRanking* and *templateSentenceScoringPuntuation* methods implement the *Template Method*.

In addition, the Factory Method design pattern[6] is used to create instances of sentence scoring services. In the present case such technique is employed to instantiate any service. The user does not need to know which class implements the service. An instance is created by request to the factory using the method name. For example, to create an instance of the WordFrequency class, it is necessary to request it to the factory by using the string "word frequency".

III. News, Blogs and Articles

Three different contexts were used to assess sentence scoring methods. The combination of sentence scoring algorithms in [3] were used in order to yield better quality results in summaries. This section describes: (i) the datasets used; (ii) the methodology followed in the assessment experiments, (iii) the abbreviations used to better understand the experiments; (iv) the results; and (v) the conclusions.

A. Corpus

Three different datasets were used in the assessment presented. They are detailed in the following subsections.

1) CNN Dataset: The CNN corpus [7] is based on the news articles in the CNN website (www.cnn.com) encompassing different subjects such as Latin America, Middle East, Europe, Travel, Business, etc. The main advantage of the CNN corpus is the high quality of the texts and the *highlights*, a good quality short summary of 3 or 4 sentences. The highlights are most important for evaluation purposes, since each of them can be seen as a gold standard summary written by the editor himself. Besides that a new evaluation test summary was developed selected by a number of researchers from the sentences of the text itself, which were the most suitable for the formation of an abstract quality, with one more option summary for evaluation purposes. The corpus CNN covers 400 texts assigned to 11 categories: Africa, Asia, business, Europe, Latin America, Middle East, US, sports, tech, travel, and world.

2) Blog Summarization Dataset: In mid-2008, Hu and colleagues [8] felt the need to organize a Blog dataset benchmark. Thus they decided to collect data from two blogs that have large numbers of posts and comments: Cosmic Variance (http://cosmicvariance.com) and Internet Explorer Blog (http://blogs.msdn.com/ie/). The evaluation dataset encompassed 50 randomly chosen posts from each blog. The reference summaries used for evaluation were generated by four people that read all the chosen posts and their corresponding comments and then labeled approximately 7 sentences from each post.

3) SUMMAC Dataset: The SUMMAC Dataset Corpus was elaborated under the responsibility of the MITRE Corporation in cooperation with the University of Edinburgh, as part of the SUMMAC conference organizer group (Tipster Text Summarization Evaluation Conference) effort[2]. This corpus is formed by 183 papers on Computation and Language, obtained from the repository LANL (Los Alamos National Laboratory) maintained by the Cornell University Library, which currently holds more than 800,000 electronic documents from various fields in their database. After selection, the documents were annotated in XML for section identification. The dataset is available through the link: http://www-nlpir.nist.gov/related_projects/tipster_summac/cmplg-xml.tar.gz.

B. Evaluation Methodology

This section describes the methodology followed in the experiments to assess the quality of summaries.

1) Quantitative assessment: ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [9] was used to quantitatively evaluate the summaries generated by using the different scoring methods. ROUGE is a fully automated widely used evaluator that essentially measures content similarity between system-developed summaries and the corresponding gold summaries.

The result of calculating ROUGE for the CNN dataset summaries is presented in two perspectives: (i) using the highlights of the CNN articles as the gold standards; and (ii) using the sentences that more closely match the highlights as the gold standards. In relation to Blog Summarization Dataset all summaries (this dataset contains four summaries as presented in section III-A2) are used as ROUGE input. For the SUMMAC Dataset, the article abstract as used as input to ROUGE.

2) Qualitative: The qualitative evaluation was performed in CNN and Blog Summarization Dataset. As already mentioned four people analyzed each original text and selected the sentences that they feel ought to be in the summary in each of the texts in the datasets. The qualitative evaluation is done by counting the numbers of sentences selected by the system that match the human gold standard. The SUMMAC Dataset provides only the article abstract, which does not fully match the assessment methodology adopted here.

C. Abbreviations

In order to facilitate the results presentation, Table I lists a set of abbreviations for the name the algorithms.

TABLE I. *Algorithms*

WF	Word Frequency
TFIDF	TF/IDF
UpCase	Upper Case
PropNoun	Proper Noun
WCOcurrency	Word Co-Occurrence
LexicalS	Lexical Similarity
CueP	Cue-Phrase
NumData	Numerical Data
SenLength	Sentence Length
SPosition	Sentence Position
SCentral	Sentence Centrality
ResTitle	Resemblance-Title
AggSim	Aggregate Similarity
TextRankS	TextRank Score
BushyP	Bushy Path

D. Results

For each dataset we uses the results of each algorithm performance [3] to analysis of some combinations of these algorithms. The combinations are:

- All algorithms;
- All word scoring algorithms;
- All sentence scoring algorithms;
- All graph scoring algorithms;
- All word scoring algorithms + all sentence scoring algorithms;
- All word scoring algorithms + all graph scoring algorithms;
- All sentence scoring algorithms + all graph scoring algorithms;
- Every combinations of the top 5 algorithms for each dataset.

1) Assessment Using CNN Dataset: The first evaluation performed uses CNN Dataset, which is a dataset containing news documents extract from CNN website (further details in section III-A1).

Table II presents the 10 best performance combinations. The combinations are obtained using the criteria presented in section III-D.

[2] http://www-nlpir.nist.gov/related_projects/tipster_summac/cmp_lg.html

TABLE II. *Combinations - CNN*

com01	LexicalS + ResTitle
com02	WF + TFIDF + LexicalS
com03	WF + TFIDF + SPosition1
com04	WF + LexicalS + SPosition1
com05	WF + LexicalS + ResTitle
com06	TFIDF + SPosition + ResTitle
com07	LexicalS + SPosition + ResTitle
com08	WF + TFIDF + LexicalS + SPosition1
com09	TFIDF + LexicalS + SPosition + ResTitle
com10	WF+TFIDF+LexicalS+SPosition+ResTitle

The result of calculating ROUGE for each the combinations is shown in Table III.

TABLE III. *Results of ROUGE having CNN dataset as gold standard applied to the proposed algorithms combinations*

	Recall	Precision	F-measure
com01	0.73(0.17)	0.36(0.12)	0.48(0,13)
com02	0.69(0.18)	0.40(0.12)	0.49(0.13)
com03	0.72(0.17)	0.36(0.12)	0.48(0.14)
com04	0.69(0.18)	0.39(0.12)	0.49(0.13)
com05	0.74(0.16)	0.36(0.12)	0.48(0.14)
com06	0.72(0.17)	0.37(0.12)	0.48(0.14)
com07	0.74(0.16)	0.36(0.12)	0.48(0.14)
com08	0.71(0.17)	0.37(0.12)	0.48(0.13)
com09	0.69(0.18)	0.39(0.12)	0.49(0.13)
com10	0.72(0.17)	0.37(0.12)	0.48(0.13)

Figure 1 presents the results of the qualitative evaluation. The highest scores were obtained by: com03 (621), com08 (621), and com10(628).

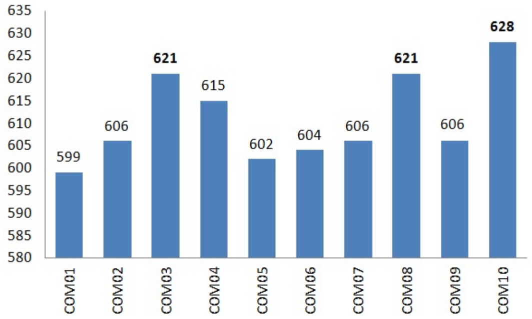

Fig. 1. Number of Correct Sentences x Combinations - Using CNN dataset

The results of this experiment are consistent as it was performed with a dataset of news, which are well structured documents. In summary:

- The documents use well-formed words, therefore the WF and tf/idf archieve good results;
- Generally, in news texts important phrases are at the beginning and at the end of a document. This explains the good results of SPosition;
- The ResTitle method archieves good results because journalists usually provide titles (headlines) containing the central information of the article;
- SCentral has good precision because this kind of texts tends to be slightly redundant;
- LexicalS archives good qualitative result because it uses synonyms to choose sentences.
- The best combinations were those that joined the best word-based and sentence-based algorithms.

2) Assessment Using Blog Summarization Dataset: The second evaluation uses a Blog dataset. The main difference from the previous experiments rests in the fact that the language used in blogs tends to be more informal and unstructured.

Table IV presents the 10 best performance combinations of the summarization algorithms for the Blog dataset.

TABLE IV. *Combinations - Blog Dataset*

com01	TFIDF + SenLength
com02	TFIDF + TextRankS
com03	WF + TFIDF + SenLength
com04	WF + TFIDF + TextRankS
com05	WF + SenLength + TextRankS
com06	TFIDF + LexicalS + TextRankS
com07	TFIDF + SenLength + TextRankS
com08	WF + TFIDF + LexicalS + SenLength
com09	WF + TFIDF + LexicalS + TextRankS
com10	TFIDF + LexicalS + SenLength + TextRankS

The result of calculating ROUGE for each the combinations is shown in Table V.

TABLE V. *Results of ROUGE having Blog Summarization dataset as gold standard applied to the proposed algorithms combinations*

	Recall	Precision	F-measure
com01	0.77(0.10)	0.63(0.14)	0.69(0,13)
com02	0.74(0.11)	0.64(0.14)	0.68(0.12)
com03	0.75(0.11)	0.62(0.14)	0.68(0.13)
com04	0.74(0.12)	0.63(0.14)	0.68(0.13)
com05	0.75(0.11)	0.63(0.14)	0.68(0.12)
com06	0.76(0.11)	0.63(0.14)	0.68(0.12)
com07	0.74(0.12)	0.63(0.14)	0.68(0.13)
com08	0.74(0.12)	0.63(0.15)	0.68(0.13)
com09	0.75(0.11)	0.63(0.14)	0.68(0.13)
com10	0.75(0.11)	0.63(0.15)	0.68(0.13)

Some points should be remarked:

- com01 and com06 achieved the best results for recall;
- com02 reached the best precision;
- com01 also achieved the best f-measure;
- com01 achieved the best results comparing other combinations and the single algorithms.

Figure 2 presents the results of the qualitative evaluation. The highest scores were obtained by: com01 (570), com07 (570), and com06(568).

Conclusions:

- Combining qualitative and quantitative evaluations the best algorithms are: com01, com06, and com07;
- The best summarization results are provided by com01 and it is the fastest in relation to algorithms that provide the best summarization results;
- com06 archives good results, but it is the slowest amongst the algorithms tested.

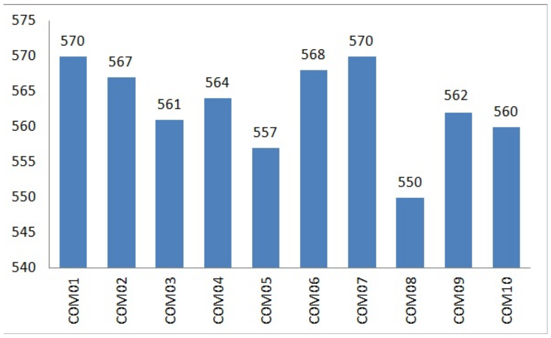

Fig. 2. Number of Correct Sentences x Combinations - Using Blog Summarization dataset

3) Assessment Using the SUMMAC Dataset: The summarization assessment of scientific papers we perform an experiment using. The SUMMAC dataset contains larger documents (in relation to other datasets used here). Each document usually is 6-8 pages long and is well structured. In order to evaluate the algorithms combinations Table VI presents the 10 best performance combinations for this dataset.

TABLE VI. *Combinations - SUMMAC*

com01	CueP + ResTitle
com02	SPosition1 + TextRankS
com03	TFIDF + CueP + SPosition1
com04	TFIDF + SPosition + ResTitle
com05	CueP + SPosition + ResTitle
com06	CueP + SPosition + TextRankS
com07	SPosition1 + ResTitle + TextRankS
com08	TFIDF + CueP + SPosition + ResTitle
com09	TFIDF + CueP + SPosition + TextRankS
com10	CueP + SPosition + ResTitle + TextRankS

The result of calculating ROUGE for each the combinations is shown in Table VII.

TABLE VII. *Results of ROUGE having SUMMAC dataset as gold standard applied to the proposed algorithms combinations*

	Recall	Precision	F-measure
com01	0.38(0.12)	0.27(0.10)	0.30(0,08)
com02	0.41(0.10)	0.26(0.10)	0.30(0.08)
com03	0.42(0.11)	0.24(0.10)	0.29(0.08)
com04	0.45(0.12)	0.24(0.10)	0.29(0.08)
com05	0.34(0.11)	0.29(0.10)	0.29(0.07)
com06	0.41(0.10)	0.26(0.10)	0.30(0.08)
com07	0.45(0.11)	0.25(0.10)	0.31(0.10)
com08	0.45(0.12)	0.24(0.10)	0.29(0.08)
com09	0.49(0.10)	0.21(0.10)	0.28(0.09)
com10	0.45(0.11)	0.25(0.10)	0.30(0.08)

Combination conclusions:

- The best results were obtained by algorithms: com05, com07, com09;
- Com09 is the best in recall, but it is the slowest one;
- Com05 is the second fastest (behind to com01) and it archived the best results for precision;
- Only com09 achieved f-measure worse than the best algorithm (alg14).

IV. General Conclusions

This paper suggests and brings experimental evidence that the effectiveness of sentence scoring methods for automatic extractive text summarization algorithms depends on the kind of text one wants to summarize, the length of documents, the kind of language used, and their structure. Different combinations of sentence scoring algorithms yield different results both in the quality of the summaries obtained and the time elapsed in generating them. The main contribution of this paper is finding the best combinations of sentence scoring methods for three kinds of documents: news, blogs, and articles.

The experiments performed allow stating that: i) Different combinations of summarization methods yield variations in the quality of summaries; (ii) The best combinations for short and well-formed texts (news) is a combination of word frequency, tf/idf, sentence position, and resemblance to the title; (iii) For blogs, short and unstructured texts, also achieve good results using word frequency and tf/idf. However, differently from news, combining these methods with text rank score and sentence length improves the results; (iv) In the case of scientific articles the best combinations include: cue-phase, sentence position, tf/idf, and resemblance to the title.

The authors are currently developing a system that analyzes the features of the input text to automatically discover which combination of sentence scoring methods best summarizes it.

V. Acknowledgements

The research results reported in this paper have been partly funded by a R&D project between Hewlett-Packard do Brazil and UFPE originated from tax exemption (IPI - Law nº 8.248, of 1991 and later updates).

References

[1] E. Lloret and M. Palomar, “Text summarisation in progress: a literature review,” *Artificial Intelligence Review*, vol. 37, no. 1, pp. 1–41, Jan. 2012.

[2] A. Nenkova and K. McKeown, “A survey of text summarization techniques,” in *Mining Text Data*. Springer, 2012, pp. 43–76.

[3] R. Ferreira, L. de Souza Cabral, R. D. Lins, G. P. e Silva, F. Freitas, G. D. Cavalcanti, R. Lima, S. J. Simske, and L. Favaro, “Assessing sentence scoring techniques for extractive text summarization,” *Expert Systems with Applications*, vol. 40, no. 14, pp. 5755 – 5764, 2013.

[4] V. G. Murdock, “Aspects of sentence retrieval,” Ph.D. dissertation, University of Massachusetts Amherst, 2006, aAI3242373.

[5] H. P. Edmundson, “New methods in automatic extracting,” *J. ACM*, vol. 16, no. 2, pp. 264–285, Apr. 1969.

[6] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design patterns: elements of reusable object-oriented software*. Addison-Wesley Professional, 1995.

[7] R. D. Lins, S. J. Simske, L. de Souza Cabral, G. de Silva, R. Lima, R. F. Mello, and L. Favaro, “A multi-tool scheme for summarizing textual documents,” in *Proc. of 11st IADIS International Conference WWW/INTERNET 2012*, July 2012, pp. 1–8.

[8] M. Hu, A. Sun, and E.-P. Lim, “Comments-oriented document summarization: understanding documents with readers’ feedback,” in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR ’08. New York, NY, USA: ACM, 2008, pp. 291–298.

[9] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, M.-F. Moens and S. Szpakowicz, Eds. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81.