

Extractive summarization: how to identify the gist of a text

Thiago Alexandre Salgueiro Pardo, Lucia Helena Machado Rino and Maria das Graças Volpe Nunes

Abstract - This paper presents a new extractive method for automatic text summarization. The new method, called gist-based method, tries to identify the gist of a text for composing its summary. GistSumm (GIST SUMMarizer), an automatic summarizer, uses such method for generating summaries, which are evaluated under the light of the gist preservation.

Index Terms – Extractive methods, gist determination, information systems, sentence extraction, text summarization

I. INTRODUCTION

Summarization is the task of abbreviating a verbal expression, textual or oral, keeping both the same relevant content and its intended effect on the reader/listener. It, thus, must concisely convey the essential original message.

People intuitively summarize texts, talks, facts and stories aiming at a big variety of purposes. Nowadays, due to the Internet and the increasing amount of on-line information, there has been a growing interest in automatic summarization, especially in text summarization. In this case, summaries of documents are very useful, since the users can rapidly select only those that are relevant for them.

Researches in automatic text summarization have been firstly tackled in the end of the 50's, when statistical techniques for extracting linguistic knowledge from texts became available. However, due to unsatisfactory results and technical bottlenecks (such as software and hardware limitations and lack of specialized knowledge for modeling summarization processes) to improve such techniques, the field of text summarization has stood still until the 80's, when computers became widely used and their components (e.g., memory and storage devices) got cheaper. At the same time, expressive linguistic resources were made available for automatic text processing.

Lately, due to the development of more powerful computers and considerable electronic resources for natural language, such as lexicons and grammars, it has been possible to fully explore more fine-grained models and theories for automatic summarization.

Although it is quite intuitive for people to summarize texts, to enable computers to do the same task is more complicated, because it involves several steps that machines should pursue, namely, finding what the main idea of the source text is and filtering what is essential in the information conveyed by the text. This step further involves differentiating complementary or superfluous information according to the intended purposes of the writers, with respect to what they aim at the readers to grasp. For example, consider the text below:

“The book *Journey to the Centre of the Earth* is a best-seller in the whole world, offering the reader the opportunity to visit a world never imagined and to enjoy fictitious adventures full of surprises that some people face when they discover a new world inside Earth. Involving and charming, this book can be found in any library near your house...”

A reader could grasp as its main idea that “the book *Journey to the Centre of the Earth* is a good book”. However, another reader could interpret it as being that “the book *Journey to the Centre of the Earth* deserves to be bought”. So, depending on the viewpoints, different information found in the very same text can play different roles. In this example, the former idea can lead to the conclusion that qualifying the book is more important than to say, for example, that “the book can be found in any library near your house”. Conversely, the latter idea can lead to the importance of expressing one's opinion, instead of giving further details of the book.

In order to tackle such a complex scenario, two main automatic text summarization approaches have arisen, namely: the deep and the superficial ones. The deep approach makes use of world and linguistic knowledge to build high-quality summaries, usually manipulating big databases and making logic inferences. Hence, the automatic summarizer should ideally simulate the human intelligence, resulting in a complex and costly process. This constrains deep summarizers to specific text genres and domains. In this case, particular features aim at reducing the complexity and the amount of information to be automatically handled. The complexity of the deep approach has motivated the superficial approach. This makes use of empirical and statistical data to determine what is

This work was supported by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico).

T.A.S. Pardo and M.G.V. Nunes are with Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Brazil (e-mails: thiago@nilc.icmc.usp.br, mdgvnune@icmc.usp.br) and are members of NILC (Núcleo Interinstitucional de Linguística Computacional), URL: <http://www.nilc.icmc.usp.br>

L.H.M. Rino is with Departamento de Computação, Universidade Federal de São Carlos, Brazil (e-mail: lucia@dc.ufscar.br) and is member of NILC

important in a text. However, although they can be widely applicable and relatively cheap, superficial methods are said to be blind, since they do not make use of linguistic knowledge to produce their summaries. As a result, these very often turn to be incoherent and, thus, useless.

As it can be seen, neither of the mentioned approaches is adequate, for both have their own bottlenecks with respect to automatic text summarization. An interesting alternative is, thus, to consider hybrid approaches, which can take advantage of their best aspects.

Letting aside the approaches, summaries can be classified as indicative, informative and evaluative. Indicative summaries are indexes to the source texts, that is, they convey only their main topics and, sometimes, do not even have a textual form. For example, they can be just a list of words. Informative summaries, on the contrary, are good substitutes of their source texts, preserving their main idea and, usually, also their structure. Differently from these, evaluative summaries add a critique of the source texts. Other distinctions can also appear between those three types of summaries: while indicative ones can be good for Internet searches, informative ones are essential for people interested in grasping most of the content of, for example, a scientific work. Additionally, evaluative ones can be important to provide enough information for the reader, to decide whether a book must be bought or not.

Summaries can still be classified according to the way they are withdrawn from the source text: they can be abstracts or extracts. Abstracts are considered to be those summaries that had been rewritten based on the original text. For example, the sentence “Dogs, cats, birds, and horses are living beings.” could be summarized by “Animals are living beings.” In this case, the rewriting of the former involved the generalization of a list of animals. Differently from using such sort of transformations, extracts involve just putting together text spans, removed directly from the text. So, abstracts are usually produced by means of deep methods, while extracts are produced through superficial ones. To stress such a difference, hereafter the latter will be referred to as extractive approaches.

We can also distinguish the authoring of the condensed texts, when considering human summarization. When it is carried out by the very same writer of the source text, it is called author (or authentic) summarizing; when the human summarizer is competent on the writing and summarization techniques, but is not necessarily a domain-expert, it is called professional summarizing.

Although deep summarization methods can produce better summaries, they are very expensive and demand very well and clearly defined linguistic and computational resources. Extractive methods, on the contrary, are cheaper and more widely applicable. So, they can present very handy solutions to the summarization problem. Having this second perspective under focus, this paper aims at introducing a new extractive method for automatic text summarization and argues that refinements of well-known extractive methods can produce useful results. Such refinements are based on the determination of the gist of the text to be summarized, which will guide the summarization process. This proposal is also interesting for

other related fields, such as Information Retrieval, Topic Detection, and Text Categorization.

Section 2 reviews some classic methods for extractive summarization, while Section 3 introduces the new method and the automatic summarizer GistSumm (GIST SUMMARizer). The method evaluation is shown in Section 4. Some conclusions are presented in Section 5.

II. SUPERFICIAL METHODS FOR AUTOMATIC TEXT SUMMARIZATION

Extractive methods of summarization comprise, basically, the following steps: (a) to identify relevant text segments; (b) to extract from the source text the minimal units (clauses, sentences or paragraphs) that contain such segments; (c) to juxtapose each one of these units to produce the final summary. Bellow, classic extractive methods are briefly described.

A. Keywords-based method

This method is based upon the fact that the writer makes use of some keywords to express his/her own main ideas [1] and these tend to be recurrent in the text. The automatic summary is, then, produced by retrieving from the source text its main keywords and putting together the minimal text units that comprise them. Some variations on this method include:

- To score each sentence according to the keywords it contains and produce the summary by grouping the sentences with the highest scores [2];
- To consider the words of the text title as keywords and produce the summary by selecting sentences that convey some of such keywords [3].

Other possibilities include, for example, considering as keywords only nouns or verbs, since these tend to be more significant in a text.

B. Localization-based method

This method assumes that the position of a sentence in a text can be associated with its importance in the context [4]. First and last sentences of a paragraph, for example, can convey its main idea and, thus, should be part of the summary.

C. Indicative and cue phrases-based method

This method selects text units with specific indicative or cue phrases, i.e., phrases considered relevant for the text being summarized [5]. For example, in scientific texts, phrases such as “the purpose of this work...”, “this paper presents...”, “results” and “conclusions” are good candidates to indicate the sentences to include in a summary. Different text genres and types can have different indicative and cue phrases. In a text about sports, for example, good phrases and words could be “the winner is...”, “championship” and “score”, instead of the former ones.

D. Relational method

This method suggests that the stronger a sentence is related to another one in a text, the more relevant it is in that context. So, more deeply related sentences should be selected to compose a summary. Additionally, the omission of some of those deeply inter-related sentences could cause non-sequitur problems, i.e., texts to be non-cohesive and, thus, incoherent [6]. The recognition of such inter-relationships usually depends upon a thesaurus, for they focus upon meaning, or the semantic account of the interconnected words and sentences.

E. Text mining-based method

Inspired on the Information Retrieval field, the idea of this method is that the more representative the words of a sentence are, the more important that sentence is. In turn, the less a word of a sentence occurs in the other sentences in the same text, the more it is representative of that sentence. So, by identifying the representativeness of words, it is possible to depict the significance of the corresponding sentences to compose the summary.

This method, called TF-ISF (Term Frequency – Inverse Sentence Frequency), was proposed in [7] and has been adapted from the Information Retrieval idea of calculating TF-IDF (Term Frequency – Inverse Document Frequency) [8] to determine how important a document is in a document collection. Instead of focusing on documents, TF-ISF focuses upon the importance of sentences in a text.

A new method, based on the Keywords and TF-ISF methods just described, is presented here. It has been defined as gist-based and has been incorporated in an automatic summarizer called GIST SUMMarizer, hereafter named GistSumm.

III. GIST SUMM: A GIST-BASED SUMMARIZER

GistSumm is an automatic summarizer based on a new extractive method based upon the gist of the source text. It tries to simulate the way human summarization occurs, in that, when a person summarizes a text, s/he first tries to identify the gist and, then, adds information drawn from the text to complement it [9]. The amount of complementary information to appear in the summary depends directly on how long the summary should be¹. In this way, GistSumm is triggered by the gist of the source text, i.e., the sentence that best conveys its gist, or the “gist sentence”. GistSumm can determine the gist through one of the following methods: the Keywords or the TF-ISF method. After pinpointing gist, it proceeds selecting sentences of the source text to appear in the summary in order to guarantee its textual features.

A. GistSumm premises

The originality of GistSumm is due to its gist determination mechanism, which guides the selection of sentences to

compose the summary based upon the indication of the gist of its corresponding source text. Such a proposal aims at improving the quality of automatic summarization, by putting together other, already well-known, techniques that have been so far used in isolation. In GistSumm, the following premises are undertaken:

- 1) Every text conveys a main idea, i.e., the gist;
- 2) It is possible to determine a sentence in the text that best expresses its gist, i.e., the gist sentence.

Based on the above, two hypotheses are adopted in GistSumm:

- Using simple statistics, it is possible to either determine the gist sentence of a text or get a quite close approximation of it;
- Knowing the gist sentence, it is possible to build coherent summaries by juxtaposing sentences related to the gist sentence. In this case, the selected sentences correlate to the gist in that they bring about complementary information.

Getting GistSumm to work, such hypotheses could be assessed, as described in Section 4.

B. Architecture of GistSumm

The architecture of GistSumm is shown in Figure 1. The following steps are executed to summarize a source text:

- 1) The source text is segmented into sentences;
- 2) Through the Keywords or the TF-ISF method (the ranking methods), sentences are ranked to determine the gist sentence. In this step, the system uses a list of stopwords (i.e., very frequent and non-relevant words), as suggested by many other superficial text processing methods;
- 3) Sentences that correlate to the gist sentence are pinpointed as candidates to compose the summary;
- 4) Only those sentences that satisfy relevance and compression rate constraints² are included in the final summary.

In what follows, sentence ranking and sentence selection processes are detailed and exemplified for the sample-text shown in Figure 2. This is a scientific text in Computer Science (already segmented in numbered sentences between square brackets).

¹ It also depends on the intended level of detail, but this is not measurable by current extractive methods.

² Compression rate is usually calculated as $1 - (\text{size of summary} / \text{size of source text})$

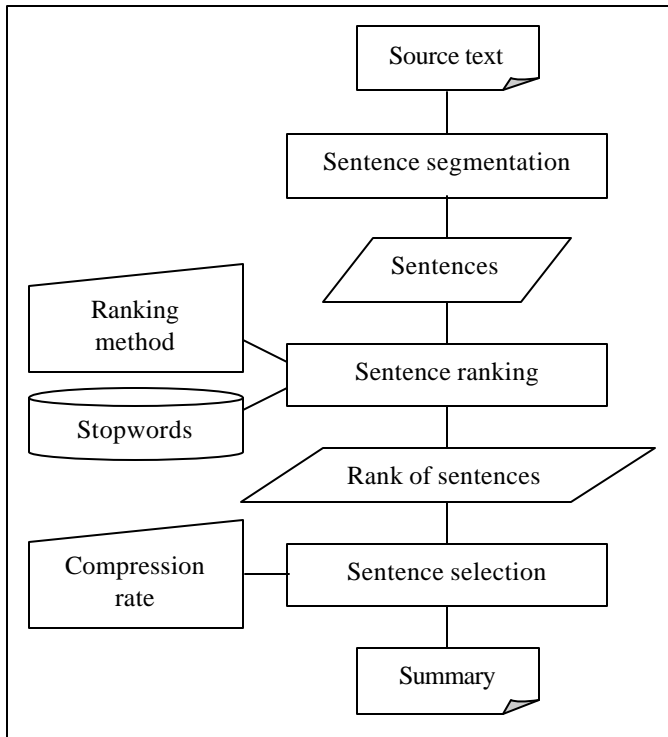


Figure 1 – Architecture of GistSumm

[English is the dominant language in the writing and publishing of scientific research in the form of scientific articles.]₁ [However, many non-natives users of English suffer the interference of their mother tongues when writing scientific papers in English.]₂ [These users face problems concerning rules of grammar and style, and/or feel unable to generate standard expressions and clauses, and the longer linguistic compositions which are conventional in this genre.]₃ [In order to ease these users' problems, we developed a learning environment for scientific writing named AMADEUS (Amiable Article Development for User Support).]₄ [AMADEUS consists of several interrelated tools - reference, support, critic and tutoring tools - and provides the context in which this dissertation is inserted.]₅ [The main goal of this research is to implement AMADEUS as an agent-based architecture with collaborative agents communicating with a special agent embodying a dynamic user model.]₆ [In order to do that we introduce the concept of adaptivity in computer systems and describe several user model shells.]₇ [We also provide details about intelligent agents which were used to implement the user model for the AMADEUS environment.]₈

Figure 2 – Sample-text

C. Sentence ranking

This process scores the sentences of the source text. This scoring happens in several steps, some of them (case folding, stemming and stopwords removal) following [10] for improving the summary generation accuracy. These steps are explained below:

- Vectoring the sentence: the sentences are represented in a word vector, keeping the original position of each word in the sentence. For example, the following vector corresponds to sentence 1 of the sample-text:

English	is	the	dominant	language
in	the	writing	and	publishing
of	scientific	research	in	the
form	of	scientific	articles	

- Case folding: each word is changed to lowercase for standardization:

english	is	the	dominant	language
in	the	writing	and	publishing
of	scientific	research	in	the
form	of	scientific	articles	

- Stemming: words with the same stem are represented only once in their first occurrence in the vector; then, the frequency of the words in the sentence are calculated and associated to the stem. In this paper, Porter's stemmer has been used [11]:

english 1	is 1	the 3	domin 1	languag 1
in 2	write 1	and 1	publish 1	of 2
scientif 2	research 1	from 1	articl 1	

- Stopwords removal: the frequency of the stopwords are reduced to zero:

english 1	is 1	the 0	domin 1	languag 1
in 0	write 1	and 0	publish 1	of 0
scientif 2	research 1	from 1	articl 1	

- Scoring the sentences: this is the most important step of sentence ranking. The scoring is carried out by one of the already defined methods, namely, the Keywords or the TF-ISF one. Using Keywords, the score of a sentence is the sum of all of its words frequencies. Using TF-ISF, the sentence score is the average of its words scores. The score of each word w is given by the following formula [7]:

$$Score(w) = F(w) \times \frac{\log n}{S(w)}$$

for $F(w)$ as the frequency of w in the sentence, n as the number of words in the sentence and $S(w)$ as the number of sentences in which w occurred.

For any of the employed methods, the sentence with the highest score is considered to be the gist sentence. Therefore, both ranking methods in GistSumm are, actually, gist determination methods.

Table 1 shows the sentences scores for both methods, Keywords and TF-ISF, when applied to the sample-text shown in Figure 2. As it can be seen, Keywords chooses sentence 4 as the gist sentence, while TF-ISF elects sentence 3. By thoroughly reading the sample-text, we can see that the Keywords method has identified more clearly its gist sentence.

Table 1 – Sentence scores

Sentence	Keywords	TF-ISF
1	24	0,465
2	22	0,628
3	23	0,671
4	42	0,598
5	22	0,643
6	37	0,663
7	17	0,571
8	25	0,575

Based on the gist sentence, GistSumm can now proceed selecting other sentences to compose the summary, as described below.

D. Sentence selection

In this process, GistSumm executes the following steps:

- 1) The average of the sentences scores is calculated to serve as the baseline for choosing the ones that will appear in the summary;
- 2) Along with the gist sentence, GistSumm selects the sentences that:
 - a. Contain, at least, one word with some of the stems of the gist sentence;
 - b. Have scores above the baseline.

Figure 3 shows the resulting summary for the sample-text, considering a compression rate of 40%, when GistSumm works on the basis of the Keywords method. Similarly, Figure 4 shows the results when TF-ISF method is used.

English is the dominant language in the writing and publishing of scientific research in the form of scientific articles. In order to ease these users' problems, we developed a learning environment for scientific writing named AMADEUS (Amiable Article Development for User Support). The main goal of this research is to implement AMADEUS as an agent-based architecture with collaborative agents communicating with a special agent embodying a dynamic user model. We also provide details about intelligent agents which were used to implement the user model for the AMADEUS environment.

Figure 3 – Summary generated with Keywords method

However, many non-natives users of English suffer the interference of their mother tongues when writing scientific papers in English. These users face problems concerning rules of grammar and style, and/or feel unable to generate standard expressions and clauses, and the longer linguistic compositions which are conventional in this genre.

Figure 4 – Summary generated with TF-ISF method

From the above, it is possible to notice that the first summary conveys the gist, while the second one does not. One can also see that the former has an unresolved reference (“these users’ problems”), whilst the latter begins with a non-contextualized discourse marker (“However”). These are typical problems resulting from extracting text spans from their context, as most extractive methods do.

Experiments have evidenced that the correct determination of the gist sentence usually produces good summaries. Therefore, the success of GistSumm depends on the success of the ranking methods for determining the gist sentence: if the method correctly identifies the gist sentence, the summary will probably be good; otherwise, the summary that fails in conveying the gist will probably not be a good summary. So, in what follows, it is described an evaluation that was carried out for trying to determine how efficient the ranking methods are for identifying the gist sentence.

IV. EVALUATING GISTSUMM SUMMARIZATION METHOD

The efficacy of GistSumm to identify the gist was evaluated by using a corpus of 10 scientific texts in Brazilian Portuguese with 530 words in average. These texts had their gist sentences selected by human judges.

Graphic 1 shows how effective GistSumm was in determining the gist sentences using the Keywords and TF-ISF methods with a compression rate of 40%. Concerning the Keywords method, the following can be verified:

- in 20% of the cases, the gist sentence was correctly identified;
- in 50%, the gist sentence got a score very close to the sentence with the highest score in the text and was selected to be in the summary;
- in 20%, the gist sentence got a score far from the sentence with the highest score in the text, but was selected to be in the summary;
- in 10%, the gist sentence got a score very far from the sentence with the highest score in the text and was not selected to be in the summary.

With TF-ISF, the following was observed:

- in 10% of the cases, the gist sentence was correctly identified;
- in 20%, the gist sentence got a score far from the sentence with the highest score in the text, but was selected to be in the summary;
- in 70%, the gist sentence got a score very far from the sentence with the highest score in the text and was not selected to be in the summary.

One can notice that the Keywords method is a better approximation for identifying the gist than TF-ISF. It is also important to notice that the inclusion of the gist sentence in the summary when it is not identified by GistSumm depends on the specified compression rate too. The higher the compression rate is, the smaller the chance is of the gist sentence being included in the summary.

Graphic 1 – Efficacy of GistSumm methods in determining the gist sentence

Method	Gist sentence identified		Proximity to gist sentence		
	Yes	No	Close	Vague	None
Keywords	20%				
		80%	50%	20%	10%
TF-ISF	10%				
		90%	0	20%	70%

V. CONCLUSIONS

This paper revisited some classic extractive methods for automatic summarization in order to propose a new one, the gist-based method. This method tries to simulate the way humans summarize texts: first, they try to determine the gist of the text and, then, they include complementary information in the summaries, in order to fulfill the needed conditions for the reader to retrieve the original gist. Following such steps, GistSumm determines the gist sentence, i.e., the sentence that best represents the main idea of the text, by means of simple statistic methods (Keywords or TF-ISF) and, based on it, selects other sentences of the source text to compose the summary.

A severe limitation of GistSumm is to rely only on just one sentence to represent the gist, as already noticed by Pardo and Rino [12]³. Considering more than one sentence to represent the gist of a text should thus be a good extension to explore. Another improvement accounts for the development of an anaphoric resolution process to minimize the lack of textuality. Such improvements, however, imply considerable extra complexity, because they usually demand deep NLP processing.

GistSumm could be used as a useful tool for applications in which the user needs to digest some amount of information in a fast way. As discussed before, a considerable tendency should be to use GistSumm for Information Retrieval. In this case, it should be possible to include other functionalities, for example, to customize GistSumm to other languages, in order to it be plugged into an Internet browser. The language customization could be done by simply substituting the stopwords list and the stemming module by corresponding ones in the newly considered natural language.

VI. REFERENCES

- [1] Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, Vol. 2, pp. 159-165.
- [2] Black, W.J. and Johnson, F.C. (1988). A Practical Evaluation of Two Rule-Based Automatic Abstraction Techniques. *Expert Systems for Information Management*, Vol. 1, No. 3. Department of Computation. University of Manchester Institute of Science and Technology.
- [3] Edmundson, H. P. (1969). New Methods in automatic extracting. *Journal of the ACM*, Vol. 16, pp. 264-285.
- [4] Baxendale, P.B. (1958). Machine-made index for technical literature – an experiment. *IBM Journal of Research and Development*, Vol. 2, pp. 354-365.
- [5] Paice, C. D. (1981). The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. *Information Retrieval Research*. Butterworth & Co. (Publishers).
- [6] Skorochoenko, E.F. (1971). Adaptive Method of Automatic Abstracting and Indexing. *Information Processing*, Vol. 2, pp. 1179-1182. North-Holland Publishing Company.
- [7] Larocca Neto, J.; Santos, A.D.; Kaestner, A.A.; Freitas, A.A. (2000). Generating Text Summaries through the Relative Importance of Topics. In M.C. Monard and J.S. Sichman (eds.), *Lecture Notes in Artificial Intelligence*, No. 1952, pp. 300-309. Springer-Verlag.
- [8] Salton, G. (1988). *Automatic Text Processing*. Reading, MA: Addison-Wesley.
- [9] Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co., Amsterdam.
- [10] Witten, I.H.; Moffat, A.; Bell, T.C. (1994). Managing Gigabytes. *Van Nostrand Reinhold*. New York.
- [11] Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, Vol. 14, No. 3.
- [12] Pardo, T.A.S. and Rino, L.H.M. (2002). DMSumm: Review and Assessment. In E. Ranchhod and N. J. Mamede (eds.), *Advances in Natural Language Processing*, pp. 263-273 (Lecture Notes in Artificial Intelligence 2389). Springer-Verlag, Germany.

³ These authors have also assumed such a limitation in their summarization system DMSumm – Discourse Modeling Summarizer.