Summarizing Scientific Texts: Experiments with Extractive Summarizers

Pedro Paulo Balage Filho, Thiago Alexandre Salgueiro Pardo, Maria das Graças Volpe Nunes
Núcleo Interinstitucional de Lingüística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
CP 668 – ICMC-USP, 13.560-970 São Carlos-SP, Brazil
http://www.nilc.icmc.usp.br
pedrobalage@gmail.com,{taspardo,gracan}@icmc.usp.br

Abstract

In this paper we present experiments on scientific text summarization. From a complete text, we produce a shorter version containing all the main parts of the research. Having in mind the sophisticated structure of such texts, we show that good results can be achieved using simple extractive summarizers with some obvious improvements that consider the specificity of the text genre. Specifically, we enhance the summarization process with the ability to detect and appropriately treat the text structure.

1. Introduction

Text summarization is the task of producing a shorter version of a text [5]. Given the incredible amount of information we have today, mainly in digital format, summarization plays an important role in everyone's life. From renting a film to selecting a book to read, people use summaries to support their decision. Summaries are also important tools for other automatic tasks. In information retrieval, for instance, it was shown that indexing summaries may be better than indexing the complete documents (see, e.g., [10]) and that summaries are useful for refining queries (e.g., [1]).

A very useful application is the summarization of scientific texts, e.g., papers and theses, either to build the summary of a text or to grasp the main ideas of it, which consist in activities that researchers have to do in their routine. Scientific texts present additional challenges to summarization in relation to "raw" texts. They present a sophisticated structure, with the text content divided in rhetorical components (in the terms of [11]), i.e., the traditional sections of a scientific text (introduction, methods, results, conclusion, references, etc.). A good summary of a text of this kind should be informative and cover the essential points of each component.

In this line, Teufel and Moens [12] made a significant contribution. They have used linguistic knowledge to classify each text sentence according to their rhetorical roles and selected the most important ones to form the summary. Such knowledge-based approach is expensive, since a rhetorical classifier must be developed on the basis of a comprehensive corpus analysis and, therefore, is highly language dependent. This linguistic data-rich approach is also supported by the work in [3], which claims that linguistic knowledge makes the difference in summarization results, a claim that is even stronger when we consider sophisticated texts, as the scientific ones.

In this paper, we present some experiments on scientific text summarization using language independent extractive summarizers. In opposition to previous work, we wanted to verify how useful such summarizers might be for such problem, since most languages do not have rhetorical classifiers and similar resources and tools available for use. We carried out experiments with a free generic extractive summarizer, the GistSumm (GIST SUMMarizer) [9], and a variation of it enriched with the ability to detect and appropriately treat the text structure. This single improvement caused the results to be better, indicating that this approach could be used for languages without refined Natural Language Processing (NLP) tools and that we did not arrive to the best of extractive summarization methods, which, in general, are linguistic-poor methods. We also argue that the incorporation of text genres specificities in the summarization methods can lead us to better results, as many other researches have already showed.

In the next section, we introduce the summarizer we used and the improvements carried out. In Section 3, we report and discuss our evaluation results. Finally, in Section 4, we make some final remarks.

2. The Summarizer

The system we based our investigation is the GistSumm [9], a language independent and generic extractive summarizer. By generic, we mean it is intended to build generic summaries, i.e., summaries for any audience; by extractive, we mean it selects and juxtaposes complete sentences from the source text to build the summary, without modifying them. We chose GistSumm for it being free, an easy-to-use system and tested for several languages and summarization requirements.

GistSumm comprises three main steps, as Figure 1 shows.

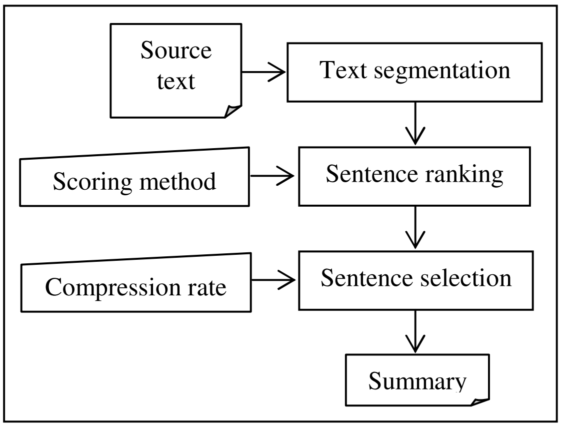

Figure 1. GistSumm architecture

The first step, text segmentation, identifies the sentences in the source text. Sentence ranking does the following: (a) stems all the words, (b) removes stopwords (i.e., too common and irrelevant words), and (c) scores each sentence in the text according to one of two scoring methods, keywords or average keywords. The first one, proposed in [2], scores each sentence as the sum of the frequency of its words in the text. The second method is a variation of the former: it simply normalizes the score of each sentence by its size, avoiding bigger sentences to be preferred in the summarization process. The highest scored sentence (by any of the scoring methods) is said to be the "gist sentence", i.e., the sentence in the text that best expresses its main idea. Finally, in the last step, sentence selection is performed. For a sentence to be selected, it must conform to two criteria: correlation with the main idea, by sharing at least one word with the gist sentence; and relevance, by having a score above a threshold, which is computed as the average of all sentences scores in the text. The number of sentences selected to be in the summary is still limited by the compression rate, which is a percentage that specifies the size of the summary in relation to the source text (computed in number of words).

In his original version, as described above, GistSumm has undergone several evaluations and showed satisfactory results. The most significant evaluation was GistSumm participation in DUC 2003 (Document Understanding Conference) [7], the main summarization evaluation conference in the area. In an usefulness task with news texts, in which each summary received a score from 0 (the summary is completely useless) to 4 (the summary is so good that could substitute the source text), GistSumm achieved the impressive average score 3,12 using the keywords scoring method. Such number was achieved for summaries formed by the gist sentences only, indicating that GistSumm selects the gist sentences with high confidence. In general, for news texts, the keywords method showed to be better than the average keywords method.

In order to summarize scientific texts, we modified GistSumm in a way that it could detect and appropriately treat the structure that these texts show. Simple heuristics were developed to identify the text sections. The heuristics look for relatively short sentences not delimited by a period, which would indicate the sections names. After determining the boundaries of each section, GistSumm individually summarizes each one, i.e., a summary (with the corresponding gist sentence) is generated for each section independently from the rest of the text. To compose the final summary, the system juxtaposes the text portions selected from each section in the source text. We will refer to this modified system as GistSumm-2.

In this system, variations in the compression rate can cause some sections to include more or less sentences than other sections in the final summary. It may also happen that some sections do not contribute to the summary. To deal with this, the system offers an option that obligates that at least one sentence from each section is included in the summary, guaranteeing that all the research parts are represented in the summary. In this case, the compression rate is automatically extended for accommodating all the sentences. This may result in the fact that the actual compression rate is not the one specified by the user.

In the next section, we evaluate the performance of both systems for scientific texts.

3. Experiments and Discussion

We carried out two experiments to verify how good the summarizers are in processing scientific texts.

Initially, we conducted a subjective evaluation, with a computational linguist judging the quality of each summary in terms of textuality and informativity. Textuality refers to coherence and cohesion in the summary; informativity refers to the amount of relevant information the summary conveys. For this experiment, we generated summaries with GistSumm and GistSumm-2 for a corpus of 20 short scientific papers on Computer Science, written in Brazilian Portuguese.

For both systems, we used the keywords scoring method, since it showed to be the best method in previous evaluations. For GistSumm-2, we used the option to include at least one sentence per section from the source text. The compression rate was set to 40% (i.e., the summary presents 60% of the size of the corresponding source text). Such a low rate was necessary to accommodate at least one sentence per section in GistSumm-2 and to allow the original GistSumm to produce summaries of similar length, so that the systems comparison could be fair. References, footnotes, and acknowledgments in the texts were removed before generating the summaries.

Table 1 and 2 show the results obtained for GistSumm and GistSumm-2, respectively, in terms of the percentage of summaries for each evaluation criterion.

Table 1. Subjective evaluation results for GistSumm

Measure	Good	Regular	Bad
Textuality	25%	55%	20%
Informativity	35%	55%	10%

Table 2. Subjective evaluation results for GistSumm-2

Measure	Good	Regular	Bad
Textuality	35%	40%	25%
Informativity	65%	30%	5%

In relation to informativity, one can see that GistSumm-2 significantly outperformed GistSumm. While GistSumm presents 35% of informative summaries, GistSumm-2 presents 65%, indicating that many regular and bad summaries produced by GistSumm had corresponding good summaries produced by GistSumm-2. In relation to textuality, the performances are practically the same for both systems. We believe this happened because textuality is a hard point to tackle that is beyond extractive summarizers capabilities.

Subjective evaluations are desirable because they show how good automatic summaries are for humans, the main consumers of such material. However, it is widely known that subjective evaluations are affected by human errors and inconsistencies, and that introduce some bias in the results. In order to avoid this, we also carried out an automatic evaluation using the traditional ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measure [4].

ROUGE is a n-gram based measure that compares an automatic summary with one or more human summaries, resulting in a score between 0 and 1. The closest the automatic summary is to the human summary, the higher the score. Because the way it works, ROUGE is basically an informativity measure, measuring, in a rough way, the amount of information the human and automatic summaries have in common.

ROUGE has been widely accepted in the research community and is one of the main automatic measures nowadays. It has been used in DUCs for more than 4 years.

For performing the evaluation, we selected the Computation and Language corpus (cmp-lg) made available by ACL (Association for Computational Linguistics) in its webpage and used in the TIPSTER Text Summarization Evaluation Conference. This corpus is composed of papers in English published in ACL conferences. We used 150 texts randomly selected from this corpus.

Each paper in the corpus was automatically parsed and had the summary, formulas, graphics and figures removed. The papers summaries were used as the human summaries required by ROUGE. Each paper had the corresponding automatic summary generated by 5 different summarization strategies:

1. by GistSumm, with the same compression rate the corresponding human summary has;
2. by GistSumm-2, with the same compression rate the corresponding human summary has and without the option to select at least one sentence per section;
3. by GistSumm-2, with the same compression rate the corresponding human summary has and with the option to select at least one sentence per section;
4. by GistSumm, with the compression rate used in the corresponding summary in strategy 3 (above);
5. by GistSumm-2, with the compression rate used in strategy 4 (above) and without the option to select at least one sentence per section.

Such strategies allow us to evaluate all possible summarization scenarios in a fair way, since that for each summary generated by GistSumm, we have one with similar length generated by GistSumm-2 with both options: at least one sentence per section or not.

We still have two more variations: considering or not the references, footnotes and acknowledgments in the text for summary production; using the keywords or the average keywords as scoring method. Here we used both methods because the evaluation is automatic and, therefore, costless.

For each summarization setup, we have run ROUGE. Here we report only the results for ROUGE-1, which basically corresponds to the co-occurrence of unigrams in automatic and human summaries. According to the experiments reported in [4], ROUGE-1 is the measure that best correlates with human judgment and, therefore, can be used alone to compare summaries. Anyway, in our experiments, ROUGE-2 to 4 and ROUGE-L (that accounts for the longest substrings co-occurrences) showed compatible results with ROUGE-1.

We report ROUGE-1 recall, precision and f-measure (with recall and precision equally weighted) in Tables 3, 4, 5 and 6. Results are showed for each summarization strategy discussed before. Table 3 and 4 show the results obtained for keywords and average keywords scoring methods for full papers. The other tables show the results obtained for both scoring methods for pre-processed papers, i.e., papers without references, footnotes and acknowledgments.

Table 3. ROUGE-1 results for keywords method and full papers

Strategy	Recall	Precision	F-measure
1	0.54626	0.16727	0.24212
2	0.35212	0.22968	0.24373
3	0.62199	0.13460	0.21142
4	0.48729	0.14827	0.20931
5	0.53799	0.16411	0.23482

Table 4. ROUGE-1 results for average keywords method and full papers

Strategy	Recall	Precision	F-measure
1	0.47655	0.15193	0.21346
2	0.47553	0.21689	0.27369
3	0.54090	0.18603	0.26013
4	0.48729	0.14827	0.20931
5	0.48906	0.21295	0.27107

Table 5. ROUGE-1 results for keywords method and pre-processed papers

Strategy	Recall	Precision	F-measure
1	0.51929	0.18088	0.25383
2	0.32319	0.25111	0.24323
3	0.59796	0.15176	0.22863
4	0.59993	0.13799	0.21204
5	0.50802	0.18768	0.25351

Table 6. ROUGE-1 results for average keywords method and pre-processed papers

Strategy	Recall	Precision	F-measure
1	0.51920	0.17662	0.24904
2	0.46155	0.23449	0.28653
3	0.51903	0.20895	0.27848
4	0.52314	0.17541	0.24604
5	0.46878	0.23166	0.28408

From the results showed, looking to the f-measure values, one can see that, in general, GistSumm-2 (strategies 2, 3 and 5) outperforms GistSumm (strategies 1 and 4), and, surprisingly, that the average keywords method is more suitable for scientific texts than the keywords method. For news texts, in previous evaluations with the original GistSumm, we had verified the opposite (see, e.g., [8]).

As expected, it is also possible to note that we have better results for pre-processed texts, since irrelevant text material was removed before summaries were produced.

We believe it is not completely fair to compare all summarization strategies at once because they have different compression rate specifications and ROUGE is sensitive to such variations. Then, making more fine-grained distinctions and comparing the most related strategies, we can realize the following:

- with only one exception (in Table 3), according to f-measure values, strategy 2 outperformed strategy 1, i.e., GistSumm-2 got better results than GistSumm for similar length summaries;
- in all cases, strategies 3 and 5 (respectively, GistSumm-2 with the option to select at least one sentence per section and GistSumm-2 without this option) outperformed strategy 4 (original GistSumm) for similar length summaries;
- in all cases, strategy 5 outperformed strategy 3, i.e., GistSumm-2 produces better results without the option to select at least one sentence per section; this indicates that GistSumm-2 is able to ignore sections that do not significantly contribute to the text overall idea.

One can also notice that, in general, GistSumm-2 produces higher precisions, balancing better recall and precision measures, mainly in strategies 2 and 5.

Undoubtedly, GistSumm-2 presents better performance than the original GistSumm, with the automatic evaluation supporting the conclusions drawn from the subjective evaluation for the informativity measure. The inclusion of the text genre specificity, namely, the detection and treatment of text structure, typical in scientific genre, has produced better results.

This shows that simple improvements in general extractive summarization methods can allow them to perform better for scientific texts, although textuality is still a point to be worked. We believe that we need more sophisticated methods to deal with it. Maybe, for this criterion, linguistic knowledge is essential, as claimed in [3].

In next section, we make some final remarks.

4. Final Remarks

Few languages have available sophisticated NLP tools and linguistic resources to perform deep linguistic processing for summarization, like the rhetorical classifier used in [12]. In this paper, we showed that a simple generic extractive summarizer may achieve better results when straightforward improvements are incorporated in the summarization process. This could be very useful for poorer languages.

More interesting, our experiments show that we still did not arrive to the best of our extractive summarizers. They can be tuned to deliver better results considering the specificities of text genres we are dealing with, as we demonstrated for scientific texts.

Independently of our results, we have no doubt that deep text understanding is essential to produce high quality summaries that, someday, may dispense human revision. As we showed, we could not improve textuality in the summaries. In fact, we believe that textuality can only be improved by using some discourse model, like Rhetorical Structure Theory [6].

5. Acknowledgments

The authors are grateful to FAPESP, CAPES, and CNPq for supporting this work.

6. References

[1] Batista Jr., W.S. and Rino, L.H.M. (2006). Extract-biased pseudo-relevance feedback. In the *Proceedings of the 4th Workshop in Information and Human Language Technology* (TIL). Ribeirão Preto-SP, Brazil.

[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, N. 3. Department of Computation. University of Manchester Institute of Science and Technology.

[3] Leite, D.S.; Rino, L.H.M.; Pardo, T.A.S.; Nunes, M.G.V. (2007). Extractive Automatic Summarization: Does more linguistic knowledge make a difference? In C. Biemann, I. Matveeva, R. Mihalcea, and D. Radev (eds.), *Proceedings of the Workshop on TextGraphs-2: Graph-Based Algorithms for Natural Language Processing*, pp.17-24. 26 April, Rochester, NY, USA.

[4] Lin, C-Y. and Hovy, E.H. (2003). Automatic Evaluation of Summaries Using N-gram Cooccurrence Statistics. In the *Proceedings of Language Technology Conference* (HLT). Edmonton, Canada.

[5] Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co. Amsterdam.

[6] Mann, W.C. and Thompson, S.A. (1987). *Rhetorical Structure Theory: A Theory of Text Organization*. Technical Report ISI/RS-87-190.

[7] Over, P. and Yen, J. (2003). An Introduction to DUC 2003: Intrinsic Evaluation of Generic News Text Summarization Systems. In the *Proceedings of Document Understanding Conference 2003*. Edmonton, Canada.

[8] Pardo, T.A.S. (2002). *GistSumm: Um Sumarizador Automático Baseado na Idéia Principal de Textos*. Technical Report. NILC-TR-02-13. São Carlos-SP, Brazil. September, 25p.

[9] Pardo, T.A.S.; Rino, L.H.M.; Nunes, M.G.V. (2003). GistSumm: A Summarization Tool Based on a New Extractive Method. In N.J. Mamede, J. Baptista, I. Trancoso, M.G.V. Nunes (eds.), *6th Workshop on Computational Processing of the Portuguese Language - Written and Spoken* (Lecture Notes in Artificial Intelligence 2721), pp. 210-218. Faro, Portugal. June 26-27.

[10] Sakai, T. and Sparck-Jones, K. (2001). Generic summaries for indexing in information retrieval. In the *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR), pp. 190-198. New Orleans, USA.

[11] Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press.

[12] Teufel, S. and Moens, M. (2002). Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, Vol. 28, N. 4, pp. 409–445.