AUTOMATIC SUMMARIZER BASED ON PRAGMATIC PROFILES

Marcus V. C. Guelpeli
Departamento de Ciência da Computação
Universidade Federal Fluminense - UFF
Rua Passo da Pátria 156 - Bloco E - 3º andar
São Domingos - Niterói - RJ CEP: 24210-240
mguelpeli@ic.uff.br
Ana Cristina Bicharra Garcia
Departamento de Ciência da Computação
Universidade Federal Fluminense - UFF
Rua Passo da Pátria 156 - Bloco E - 3º andar
São Domingos - Niterói - RJ CEP: 24210-240
bicharra@ic.uff.br

ABSTRACT

The aim of this paper is to make a proposal for an automatic summarizer based on pragmatic profiles. It uses a traditional algorithm for the field/area of automatic summarization to be compared with the results of the profile algorithm. This methodology was developed to emphasize the importance of the word within each sentence as an index of grammatical development. In this way, we propose a classification of the original text in relation to its temporal measurements and textual composition vis a vis its formality, allowing for parameters to determine the level of compression automatically in order to generate the summary.

KEYWORDS

Automatic summarizer, Pragmatic profiles, Ideal extracts, Automatic compression.

1. INTRODUCTION

According to Mani and Maybury (1999), text summarization is a process that attempts to create a shorter version of an original text. The need to simplify and summarize a text occurs due to the increase in the volume of information available in the means of communication and the short time people have to read texts on a variety of subjects. As a consequence of this process readers find themselves unable to absorb the content matter of the original texts. Hence, the summary is a short text with the aim of capturing the author's main point and, in few lines, communicating it to the reader.

Automatic Summarization (AS) has been formalized since 1950, although the initial breakthroughin the research was Luhn's method (1958) on the keyword. From that date on, other works were conducted in the field, such as:

Edmundson (1969) discusses the computational choice of sentences with the greatest potential of communicating the meaning of the original text; Maybury (1999) proposes the use of the hybrid approach and Hovey, Lin and Zhou (2005) describe the use of Basic Elements to compress sentences in the multiple documents summarization. As established in works that delimit AS, one can verify a methodological taxonomy where there is a surface approach process, also described as empirical or statistical, and another approach known as deep or fundamental approach.

The current AS's find it difficult to generate summaries that maintain a degree of faithfulness to the thoughts of the authors of the text, as well as finding it hard to adapt the summary to the readers' interest, regardless of whether they choose a deep or surface approach. This paper proposes the use of a hybrid approach, with the development of pragmatic profiles for the creation of summaries that better reflect the

authors' ideas, by means of using the style rules proposed by Hovy (1988). The main motivation would be the use of this summarization technique for reading texts from the internet, where there is a very large volume of information.

2. LANGUAGE ACQUISITION AND THE DEVELOPMENT OF WRITING

In the summarization process, sentences are generated by way of a selection of words. The use of the word within each sentence – as an index of grammatical development and as a metric for the composition of the summary – has as its reference the theoretical foundation proposed by Brown (1973), who states that the best indicator of languages adherence is the Mean Length of Utterance (or MLU) by way of the morphemes (the average number of morphemes used in the vocalizations). The MLU is an effective indicator to measure grammatical evolution. As the number of words of an individual increases, this indicator follows it, reflecting the evolution and increase in the length of the text.

Magalhães (2006), uses this metric in her methodology to assess that individuals in the same age group have differentiated syntactic rates. This metric is different from the one used by Brown (1973), basically due to the fact that it considers the number of words in sentences rather than the notation of a morpheme.

3. THE ACQUISITION OF A PRAGMATIC PROFILE

In the deep approach, Hovy (1988) proposes the use of pragmatic profiles, where certain metrics are established, which the author calls style characteristics. He establishes a temporal relation in the preparation of the original text and classifies it as scarce, little, sufficient or unlimited, as per Table 1 (below). The author also addresses the type of textual writing, in which he bases the use of certain rules of classifying texts according to their formality, such as colloquial, normal or formal.

Table 1. Represents the interaction between style rules and implications in the content of the summary (Hovy, 1988).

Time **Formality**	**Scarce**	**Little**	**Sufficient**	**Unlimited**
Colloquial	High summarization; only the main topic	Medium summarization	Medium summarization; main topic, unknown details	Low summarization; main topic, relevant details
Normal	Medium summarization; main topic, a few details	Medium summarization; main topic	Medium summarization; main topic, important details	Low summarization; main topic, relevant details
Formal	Medium summarization; main topic.	Medium summarization; main topic	Low summarization; main topic, important details	Low summarization; main topic and related, relevant details

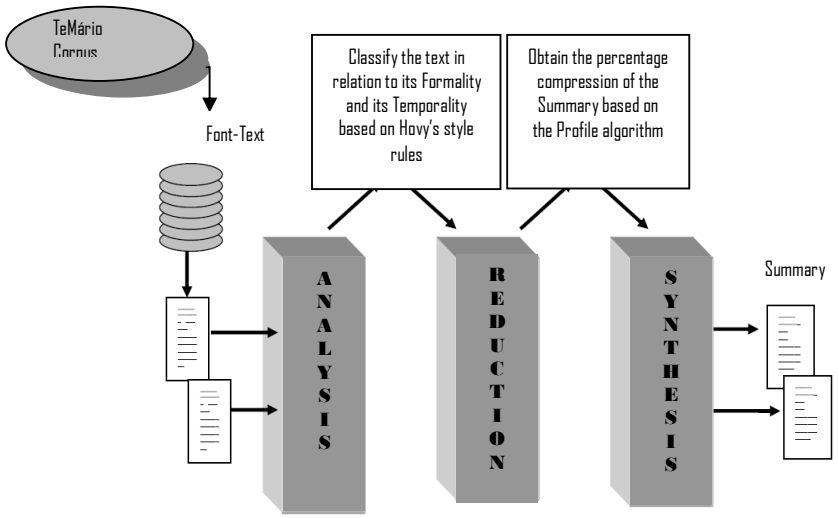
The texts with formal characteristics tend to have longer phrases; that is, they have a greater number of words. On the other hand, more colloquial texts tend to have lower number of words. This is attested by studies in language acquisition and the development of writing. There is a guarantee that the size of the summary should be coherent according to the degree of formality of the text, consequently determining the compression size of the summary (Hovy, 1988).

Another extremely relevant characteristic is the temporality in the construction of the texts. This metric is also important because it is another determining factor in the composition of the summary. The less time one has, the less should be included in the summary, be it due to a spatial limitation or due to the knowledge level of the reader (Hovy, 1988).

4. METHODOLOGY

The methodology for the automatic summarization adopted in this work has a hybrid approach (a surface and deep approach), as illustrated in Figure 1. The proposed summarizer uses the extraction and transposition of sentences for the composition of the summaries, respecting their position in the original text (a feature of the surface approach), while also adopting the deep approach when it uses Hovy's rules (1988) to classify the original text, basing itself on the user's pragmatic profile. This classification is determined by the algorithm proposed in this work, called Profile, which will be described in detail in this section.

Fig 1. Functionality of the Automatic Summarizer structure using the Profile algorithm.



This work compared the TextTiling algorithm used by Hearst(1997) and Larroca et al (2000), which initially allows the division of a text into the various segments of which it is composed. The *Term Frequency–Inverse Document Frequency* (TF-IDF) is calculated, followed by the relevance average of each term in the text by way of the *Term Frequency–Inverse Sentence Frequency* (TF-ISF).

The linguistic corpus used for the summarization in this work was the TeMário Corpus, proposed by Pardo and Rino (2003). The idea of the Ideal Extracts Generator for Brazilian Portuguese by Pardo and Rino (Pardo and Rino 2004) was also used. These results worked as parameters for the comparison of the Profile algorithm.

4.1. How the Automatic Summarizer Works

In Figure 1, the summarizer uses the font-text obtained from TeMário Corpus, in which taxonomy is used according to its formality and temporality (analysis phase) based on style rules proposed by Hovey (1988). With that, the Profile algorithm is applied to determine the degree of compression that will be used to obtain the summary, reflecting the author's pragmatic profile (reduction phase), without there being any human interference. The summary carries out the extraction and transposition of sentences, respecting its position in the original text, making up the profile with the sentences with the highest frequency of words determined by the profile algorithm (synthesis phase).Another relevant fact about the proposed summarizer it isen't use of *stopwords* used in the analysis phase.

5. EXPERIMENTS

To test the strategies of the automatic summarization of the text, an environment was created using Visual C++ 6. The Automatic Summarizer was implemented to work with the TextTiling and *Profile* algorithms.

5.1. Evaluation of the Implemented Strategies

Evaluations of automatic summaries are rather difficult and involve human evaluation, making them quite costly. Teufel and Moens (1999) suggest adopting ideal extracts for the automatic evaluation of generated summaries. An ideal extract should contain sentences that correspond to the content of the manual summary. Ideal extracts allows for the use of new methodologies of automatic summarization and they are built or indicated by readers or competent authors, thereby representing the best reference for the evaluation of AS systems (Parde and Rino, 2004). This work chose the TeMário Corpus created by Pardo and Rino (2003) to conduct the tests of the summaries and the results of the Ideal Extracts Generator for Brazilian Portuguese, also by Pardo and Rino (2004). This corpus is made up of 100 texts, which are classified by Summaries and Source Texts. In the conduction of the experiments, we used the source text with the origins and title. In the division used in the simulation, there is a subdivision with texts from two major Brazilian newspapers: Folha de São Paulo and Jornal do Brasil. Ideal automatic extracts were used in the summaries to compare the results obtained with the proposed summarizer using the TextTiling and Profile algorithms.

5.2. Obtained Results

The results were obtained used Pardo and Rino's TeMário Corpus (2003), presenting the behaviour of the TextTiling algorithm and the Profile algorithm. These values were measured based on the ideal extract. It used the measurements of an intrinsic evaluation, which considers the aspects of content matter and quality (Pardo and Rino, 2004): *Recall*(R): number of sentences from the automatic summary that are present in the reference summary / number of sentences in the automatic summary; *Precision*(P): number of sentences in the automatic summary that are present in the reference summary / number of sentences in the reference summary. *F-Measure*: ((P x R)/(P + R)) x 2; the closer this is to 1, the better the summary.

Another measure used in the evaluation was the compression and the number of sentences in the original text that were maintained in the summary, whereby the compression is expressed by the number of sentences eliminated in order to make up the summary and another measure would be the sentences from the original text that were maintained in the summary. In compression, the closer the value is to 0, the larger the summary will be; that is, a high number of sentences from the original text will be maintained and, conversely, in values closer to 1, the opposite is true.

Table 2 represents the average results of the metric applications for all the sections of the Corpus. In the last item, percentage of sentences in the original text that were maintained, the results with a lower percentage are more relevant because a more compact summary was created. Note that the results obtained in the Profile algorithm are, on average, better for all sections and using any metric, in comparison to those obtained with the TextTiling algorithm.

Table 2. Average Result of the application of the TextTiling and Profile Algorithms in the corpus with all measures.

News	Sections	Recall		Precision		% Compression		% Sentences in the original text that were maintained	
		Text Tiling	Perfil	Text Tiling	Perfil	Text Tiling	Perfil	Text Tiling	Perfil
Folha SP	World	24.43	31.89	33.33	43.93	61.65	64.00	38.35	36.00
	Opinion	23.62	35.80	28.78	37.87	60.40	66.40	39.60	34.00
	Special	25.72	37.72	33.45	44.11	59.90	63.80	40.10	36.20
JB	International	36.16	52.90	38.91	46.92	59.05	67.20	40.95	32.80
	Politics	25.30	40.07	33.61	47.18	59.35	64.90	40.65	34.65

6. CONCLUSION

In this work, a proposal was made for an algorithm called Profile based on Hovy's rules (1988). The results obtained are encouraging, seeing as they were favourable when compared to the TextTiling algorithm, which is widespread in the literature.Using pragmatic profiles in text summarization may offer a number of advantages, such as: personalising summaries according to each author's pragmatic profile, determining the level of compression automatically based on each author's profile as well as guaranteeing a better quality in the summaries, as it warrants a greater degree of faithfulness to the author's idea. This work can still be perfected in regards to the parameters of the percentage compression, which are fixed, seeing as these can be learned as the result of each author's interactions.

The results obtained herein can be adapted for Automonous Learning using the Hidden Markov Models (HMM) concept and profiles can be created according to Guelpeli et al. (2004) using learning by reinforcement for the autonomous modelling of the author as well as the reader. This idea could be extended and used on the Internet as a way of summarizing news that the user tends to read often.

REFERENCES

Brown, R, (1973).*A first language.* UNIVERSITY Press, Cambridge. http://bowland-files.lancs.ac.uk/chimp/langac/LECTURE2/2brown.htm, acessado em dezembro de 2006.

Edmundson, H. P. (1969).*New Methods in automatic extracting.* Journal of the ACM,16, pp. 264-285.

Guelpeli, M., Ribeiro, C., and Omar, N. (2004) *Aprendizado por Reforço para um Sistema Tutor Inteligente sem Modelo Explícito do Aprendiz*, Revista Brasileira de Informática na Educação- RBIE – SBC Volume 12 – Número 2 pág. 69-77 - mês de Julho a Dezembro de 2004 ISSN 1414-5685.

Hearst, M. A.(1997).*TextTiling: Segmenting text into multi–paragraph subtopic passages.* Computational Linguistics, vol. 23, no. 1 pp. 33–64, 1997. Disponível em: http://ucrel.lancs.ac.uk/acl/J/J97/J97-1003.pdf , acessado em 06 de Maio de 2007.

Hovy, E.H, Lin, C.Y. and Zhou, L. (2005). *A BE-based Multi-document Summarizer with Sentence Compression*. Proceedings of Multilingual Summarization Evaluation (ACL 2005 workshop). Ann Arbor, MI

Hovy, E. (1988).*Generating Natural Language under Pragmatic Constraints.* Lawrence Erlbaum Associates Publishers, Hillsdale, 224 page ISBN 0805802487.

Larocca, J. N.1, Santos,A. D. S, Kaestner ,C. A.A. e Freitas A. A. (2000). *Generating Text Summaries through the Relative Importance of Topics.* Lecture Notes in Computer Science Springer Berlin / Heidelberg Volume 1952/2000, ISSN0302-9743 (Print) 1611-3349 (Online) pp 300, 2000, Brazil.

Luhn, H. P. (1958). *The automatic creation of literature abstracts.*IBM Journal of Research and Development, 2, pp. 159-165.

Magalhães, T.M.V.(2006).*O Sistema Pronominal Sujeito e Objeto na Aquisição do Português Europeu e do Português Brasileiro.* Tese de Doutorado, UNICAMP.

Mani, I.; Maybury, M.T. (1999). *Advances in automatic text summarization.*MIT Press, Cambridge, MA.

Pardo, T.A.S. e Rino, L.H.M.; Martins(2003).*TeMário: Um Corpus para Sumarização Automática de Textos*.Relatório Técnico número NILC-TR-03-09, Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional, NILC - ICMC-USP, Outubro de 2003.Disponível em http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm acessado em 20 de Abril de 2007.

Pardo, T.A.S. e Rino, L.H.M.; Martins(2004).*Descrição do GEI – Gerador de Extratos Ideais para o Português do Brasil.*Relatório Técnico número NILC-TR-04-07, Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional, NILC - ICMC-USP, Agosto de 2004.

Teufel, S. and Moens, M. (1999). *Argumentative Classification of Extracted Sentences as a First Step Towards Flexible Abstracting.* In I. Mani and M.T. Maybury (eds.), Advances in Automatic Text Summarization, pp. 155-175. MIT Press, Cambridge, MA.