A Methodology for Noun Phrase-Based Automatic Indexing

Renato Rocha Souza* and K.S. Raghavan**

* Professor, Departamento de Organização e Tratamento da Informação, School of Information Science, Federal University of Minas Gerais, Av. Antônio Carlos 6627, Belo Horizonte, MG, Brazil 31161-970 <rsouza@eci.ufmg.br>

** K. S. Raghavan, Professor, Documentation Research & Training Centre, Indian Statistical Institute, Mysore Road, Bangalore 560 059, India < raghavan@drtc.isibang.ac.in>

Renato Rocha Souza is a Professor at the School of Information Science, Federal University of Minas Gerais (UFMG); He is a Sub-coordinator of the School's Post-Graduate program in Information Science and is also the Vice-chief of its Department of Organization and Handling of Information; His research interests include: classification and categorization, automatic indexing, and epistemology of Information Science.

K. S. Raghavan is a professor at DRTC, Indian Statistical Institute, Bangalore, India. Prior to joining DRTC, he was Dean (Academic) and Professor & Head of the Department of Information Science, University of Madras. He is a Corresponding Member of the IFLA Section on Classification & Indexing. He served as a visiting professor at the School of Information Science, Federal University of Minas Gerais (UFMG) in 2003. His research interests include: Knowledge Organization in the Digital Environment, Multilingual Thesauri, Lateral Semantic Relations and Scientometrics.

Rocha Souza, Renato and K.S. Raghavan. **A methodology for noun phrase-based automatic indexing.** *Knowledge Organization, 33(1)* 45-56. 28 refs.

ABSTRACT: The scholarly community is increasingly employing the Web both for publication of scholarly output and for locating and accessing relevant scholarly literature. Organization of this vast body of digital information assumes significance in this context. The sheer volume of digital information to be handled makes traditional indexing and knowledge representation strategies ineffective and impractical. It is, therefore, worth exploring new approaches. An approach being discussed considers the intrinsic semantics of texts of documents. Based on the hypothesis that noun phrases in a text are semantically rich in terms of their ability to represent the subject content of the document, this approach seeks to identify and extract noun phrases instead of single keywords, and use them as descriptors. This paper presents a methodology that has been developed for extracting noun phrases from Portuguese texts. The results of an experiment carried out to test the adequacy of the methodology are also presented.

1. Introduction

Information and communication technologies have revolutionized the ways in which individuals, scientific communities and communities of practice disseminate, exchange, access and use information. A consequence of the increasing use of the Web to publish information and the resulting emergence of the Web as a major source of information is the recognition of the need for better organization of the infor-

mation on the Web to facilitate more effective retrieval than what is possible using the available search engines. Developments and initiatives such as metadata initiatives (e.g. DCMI), Semantic Web, ontologies, tools and technologies for data and text mining, etc. should all be viewed against this background.

That there has been a significant increase in the volume of scholarly information that is being published and disseminated in electronic form via the Web is now a widely recognized fact. Managers of information and knowledge in organizations are increasingly facing a situation of both information overload on the one hand and increasing demand for filtered and relevant information on the other. Considering the sheer volume of information to be handled, it is widely recognized that solutions to handle such a situation should necessarily be technology-based and make effective use of intelligent technologies. When one looks at the history of developments in information retrieval and IR systems, it becomes obvious that this is not something entirely new. Information retrieval systems have, in the past, experimented with different strategies and new technologies. H.P. Luhn's experiments in keyword indexing and Selective Dissemination of Information, in the middle of the 20th Century, are among the early efforts – and such efforts have continued to this day – directed at more efficient handling of information to facilitate and enhance retrieval and dissemination. Strategies that utilize digital computer technologies to manage large document collections have been in use for sometime now. Intranets including corporate portals, subject gateways and digital libraries are all developments along these lines.

An improvement in the effectiveness and efficiency of such strategies naturally depends on research and development in several areas. The major approaches and strategies that have been adopted in the experiments aimed at enhancing retrieval effectiveness could be grouped into two broad categories:

- Strategies aimed at improving the quality of metadata extracted from and/or assigned to resources
- Strategies aimed at enhancing search interface to facilitate more effective retrieval through meaningful navigation.

These are not entirely independent of one another and, in practice it is possible to adopt a combination of the two strategies in the design of information systems. An overview of the principal research directions that are being pursued is given below:

- Exploring the possibility and feasibility of utilizing the semantic and semiotic information intrinsic to the text (or other kinds of media) to arrive at more meaningful and accurate representations of the 'aboutness' of the document; indeed, a very broad range of tools and approaches are being experimented with;
- Adopting and implementing appropriate metadata standards at the 'pre-publication' stage of a digital resource (a kind of 'prenatal cataloguing', as S. R. Ranganathan put it or 'cataloging at source/in publication' as Library of Congress would have it) to semantically mark the data for easy identification and manipulation by computers, search engines, etc.;
- Developing more sophisticated tools and technologies (Computational Linguistics, Natural Language Processing, etc) to process texts and extract valuable and useful metadata to serve as search keys; these aim at deriving more significant and acceptable representations to index texts and to serve as points of access to improve retrieval effectiveness;
- Developing and adopting such tools for knowledge organization as ontologies, concept maps, etc.;
- Cognitive approaches to information retrieval involving a range of strategies such as creation of user profiles and building intelligent IRS that *learn* from the interactions with the user and utilize this knowledge to enhance retrieval; and,
- Value additions in the display and presentation of retrieved information such as presenting contextual information to minimize 'noise' in retrieval, providing hyperlinks to retrieved and related documents, etc.

Probably many other strategies in the context of multimedia and image files are being experimented with. It should be apparent from the above that a comprehensive approach to enhancing digital information organization and retrieval should bring together these strategies based on inputs from several areas including Information Science, Linguistics, Computer Science, Cognitive Psychology, Communication, etc. and integrate these to find workable solutions to the problems of information representation, organization and retrieval. (The terms 'information representation' and 'knowledge representation' will be used interchangeably in this paper.)

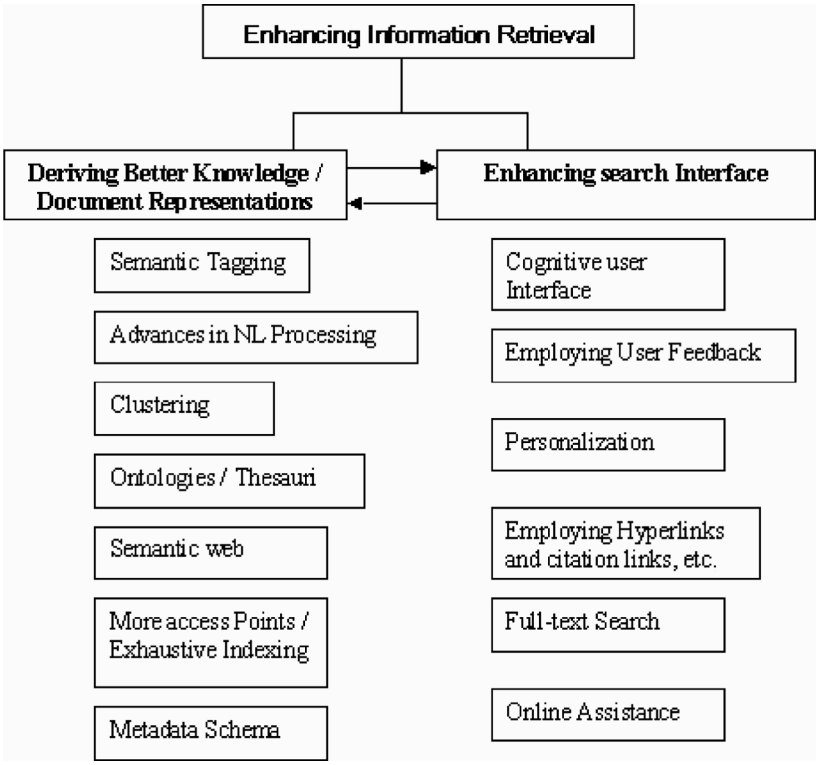

Figure 1. Strategies for Enhancing IR

2. Noun Phrases and Information Retrieval Systems

This research focuses on improving knowledge representations derived from texts with a view to enhancing information retrieval. There are suggestions to the effect that identification and extraction of noun phrases (NPs), instead of keywords, may prove to be a useful strategy for selection of index terms. This strategy is based on the hypothesis that NPs carry the greater part of the semantics of a document, as opposed to articles, verbs, adjectives, adverbs and connectives (Baeza-Yates & Ribeiro-Neto 1999, 169-70). For example, the NP 'information science' conveys more information than a Boolean combination of the words 'science' and 'information'. NPs are parts of a larger text structure, and bring about a degree of cohesion between the components of the larger structure (Perini 1995). Le Guern and Bouché (quoted in Kuramoto 1999) view a noun phrase as the smallest unit of information contained in a text. The SYDO research group, to which these researchers belong, has, as one of its objectives, the utilization of NPs as descriptors. Identification of NPs generally employs syntactic distance (measured by the number of words in between) within a predefined limit as the criteria. Use of NPs in information retrieval systems and their use as descriptors is, in effect, an extrapolation of this.

Parts-of-speech tagging has been used for English since the 1960s, and has reached a fairly acceptable level of accuracy. Some work on syntactic marking of Portuguese language texts and developing tools for automation of noun phrase extraction from these texts has been reported from the Southern Denmark University (Bick 2000, Vieira 2000, and Projeto Dirpi 2001). Noun phrase extraction has a wide range of applications including indexing and information retrieval. Extraction of NPs has been found to be useful for translation of concept maps (Woods, Richardson & Fox 2005) and even in automatic translation. Suggestions for employing linguistic approaches in information retrieval are also not new. As early as in 1983 Salton & Mcgill (1983, 90-94) had discussed the use of linguistic methods in information retrieval. Of course, they elaborate with illustrations, the difficulty and near impossibility of unambiguously and accurately recognizing semantics of a document through analysis of components of a sentence. They suggest that a model based on transformational grammar could lead to better results. The authors are, thus, in agreement with Liberato (1997) who contends that adequate and accurate semantic analysis of texts is possible only through con-

textual cognitive analysis. Certain proposals including examination of the efficacy of phrase-generating algorithms based on frequency of words and human interference in the disambiguation process to overcome this problem have also been made.

Information retrieval systems usually adopt keywords for indexing. It is often contended that the semantics of the texts of documents and user needs (e.g. as in a query) can be expressed through Boolean combination of single words. This is clearly an over simplification of the actual problem as a great part of the semantics of the document, or the user query, is lost when the text is represented by a Boolean combination of words (Baeza-Yates & Ribeiro-Neto 1999, 19). Some of the works that specifically look at the value and utility of 'noun phrase extraction' in the information retrieval context are those by Kuramoto (1996 and 1999), Moreiro et al (2003), and Velumani & Raghavan, (2005 & 2006). Kuramoto explores the potential of NPs as descriptors of value in information retrieval. Velumani & Raghavan report on the utility of employing a combination of available online validation tools (such as online glossaries, online thesauri, etc.) and frequency data for identifying and extracting '*content rich*' NPs from HTML texts. The work of Kuramoto has been the principal influence for the present work. However, in Kuramoto's research (1996, 6): "the extraction of NPs was done manually simulating automatic extraction. This procedure was adopted primarily because of the lack of a system for automatic extraction of NPs in collections containing documents in Portuguese". Today, there is at least one tool that is available for such work (Gasperin et al., 2003) and it was thought that it is worth examining its application and utility. Another fundamental difference between the work reported in this paper and that by Kuramoto is that while Kuramoto focused almost entirely on IRS based on NPs, this work is aimed at developing a methodology to aid automatic indexing and derivation of representations by processing texts.

The syntagms that constitute a sentence may or may not be easily recognizable. It may be necessary to use other resources. Perini (1985, 42-43) believes that the intuitive processes employed by humans can be formalized and that it is possible to build algorithms to identify and extract portions such as NPs from texts. Probably many others from the computational linguistics field share this view. However, there are those (especially from the field of Linguistics) who hold the view that, given the ambiguities of natural languages and language usage by individuals, complete and accurate identification of constituents of a text requires a cognitive and contextual approach and is possible only through human efforts (Liberato 1997). There are also linguistic models (e.g. Noam Chomsky's *Transformational Model*), which are far more difficult for reduction to algorithm-based procedures. (Ruwet 1975, 155-212, 223-79) No doubt, all the views mentioned above are valid within their own contexts. What is of importance in the context of this paper is the fact that machines can be, and have been, programmed to understand the structure of sentences and successfully extract required portions at an acceptable level of accuracy. Of course there are problem situations and there are limits to this capability of computers. But the question is not whether there will be errors in machine-based text processing for noun phrase extraction; it should be obvious that there will be. But if such errors could be kept within acceptable limits (and this should be especially possible for scientific texts) and the *noise* that can occur as a result of this within tolerable limits, it is worthwhile to explore the feasibility and utility of developing mechanisms for automatic extraction of NPs to enhance information retrieval and for other possible applications.

According to Miorelli (2001), NPs can be understood and treated syntactically focusing on the form, or semantically, looking for greater meanings. The pragmatic-semantic approach employed by Liberato (1997) requires a 'context interpreter', which is natural to human cognition but not adequately implemented in artificial intelligence heuristics. Another issue of importance in the context of extraction of NPs from texts is that related to the extensive use of anaphors (e.g. pronouns) in natural language texts as substitutes for an entity, object or event (Vieira 1998 and 2000; Sant'anna 2000; Rossi et al. 2001, Gasperin et al. 2003). Understanding these is useful in deriving semantic representations of a text and it can substantially improve the quality of results in several applications of natural language processing including indexing, information retrieval and even automatic translation (Rossi et al. 2001). Anaphoric resolution has to be necessarily based on some kind of context analysis, even if computationally done. The syntactic form, as analyzed by Perini (1986, 1995 and 1996), is more related to the structures of the sentences per se, and is more easily handled by computers. Miorelli's broad approach (2001) is widely used in experiments related to automated extraction of NPs and the research reported in this paper also employs this approach.

3. Objectives

As should be evident from the foregoing review, an important question in automatic indexing is how to extract semantically rich terms from a text. Semantic richness is used here to mean the ability of a term or phrase extracted from a text to accurately and meaningfully represent full or partial subject content of the text. References have already been made to the several research studies that have contributed to a better understanding of the many issues that emerge when dealing with massive amounts of natural language texts that need to be organized for effective search and retrieval as in large collections of digital documents. Our research seeks to carry forward utilization of semantics embedded in texts in Portuguese language for deriving meaningful representations of their 'aboutness'. We explore the potential use of NPs as descriptors of documents because of the higher degree of semantic information they contain.

The research reported in this paper is based on the following hypotheses:

– NPs in a text are semantically richer and thus constitute better metadata for representing the 'aboutness' of documents than mere keywords or other portions of the text.
– It is feasible to develop and implement effective mechanisms for discovering and extracting content-bearing NPs from texts to create searchable and browsable indexes of full-texts.

A methodology for automatic identification of NPs as descriptors, instead of keywords, is proposed and tested with a sample corpus of digital texts in Portuguese.

4. Methodology

There are two principal aspects to this paper. The first relates to the methodology for automatic extraction of NPs from texts. The second aspect relates to the methodology for computing and assigning a weight (score) to every extracted NP indicating its utility and value as a descriptor vis-à-vis the other NPs extracted from the same text.

4.1 Noun Phrase Extraction

Figure 2 gives a step-by-step description of the general methodology for noun phrase extraction adopted in this research. Once a corpus of texts in a domain is chosen, the system requires that all the documents be converted into simple text files for

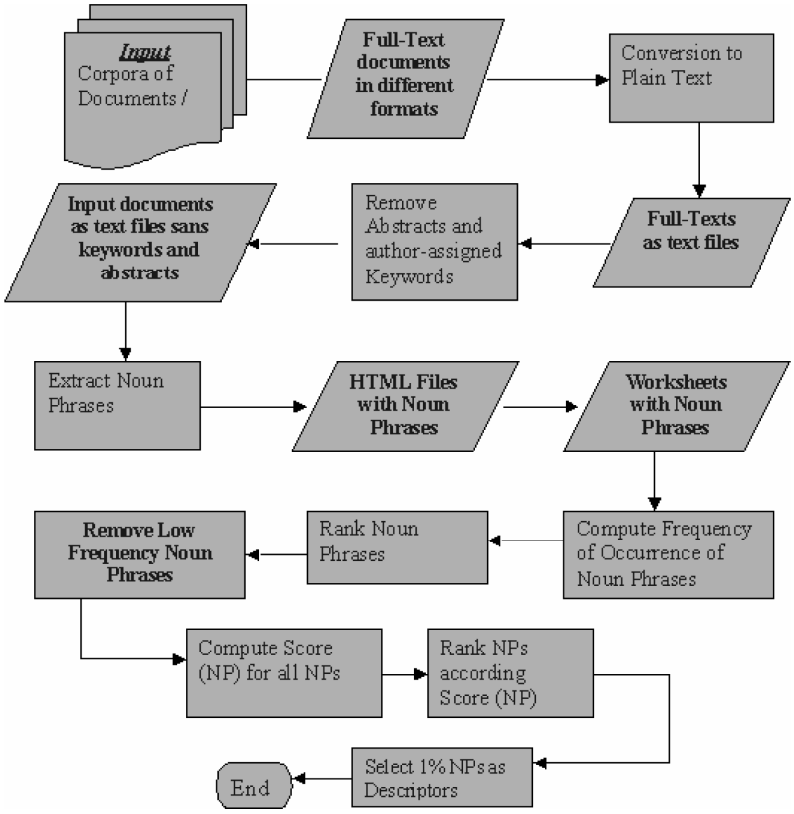

Figure 2. Noun Phrase Extraction Proces

further analysis and extraction of NPs. A few comments by way of explanation of the procedure are in order. While the methodology can be employed to compute the score for all the NPs extracted from a text, it is better to eliminate very low frequency NPs before computing the Scores for NPs. First of all very low frequency NPs are not likely to be quite useful as descriptors. Secondly in the actual experiments conducted it was found that low frequnency NPs constituted nearly 80% of all NPs extracted. It may also be mentioned here that the methodology adopted here requires that for every NP extracted from any text, its occurrence in every other text in the corpus also needs to be computed. Thus eliminating them at this stage saves a considerable amount of computational effort.

4.2 Assigning weights to NPs

Once a corpus of texts in a domain is chosen, the system requires that all the documents be converted into simple text files for further analysis and extraction of NPs. Computing a score indicating the value and utility of a NP as a descriptor in the context of a given text / document requires the identification and adoption of a valid criteria for assigning such a score. This should necessarily be based on an understanding of the factors that determine the importance of the NP in a given text. Word frequency has for long been used as a valid criterion in determining the utility and value of a word in a text. This is based on the hypothesis that high frequency words are more likely to be useful in representing the semantics of a text than low frequency words. Therefore in this research one of the principal criteria used in assigning a weight to a NP is the frequency of its occurrence in the concerned text. Another important factor that should be considered in assigning a score to a NP is its distribution among the texts in the corpus/domain. For example, a NP that is more evenly distributed among the texts in a domain/corpus (i.e. common to a large number of documents in the corpus) has a very low discriminating value and is less useful as a descriptor and as a search key. As against this a NP that is unique to one or only a few documents is likely to be more useful and acceptable as a descriptor for the document(s). The third factor that has been taken into consideration in this research for arriving at a score is its position within a sentence in the text. The methodology adopted here takes all these three factors into consideration in arriving at a score for a NP. A brief explanation of the procedure developed for computing the scores of NPs is given below. Every noun phrase is assigned a score computed using the following formula:

$$Score\ (NP) = [(k1 * Tf\ (X)) - (k2 * IDf\ (Y)) + (k3 * CNP)]$$

Where:

Score (NP) is a weight computed for a NP indicating its utility and value as a descriptor to represent the '*aboutness*' of the source document.

Tf (X) = frequency of occurrence of the NPs in the document after correcting for distortions;

IDf(Y) = the number of documents in which the NPs occurs with frequency higher than Y; This factor reduces the weight assigned of a NP and its Score (NP).

CNP = another value assigned to a NP depending on the category to which it belongs.

Some explanation of the values assigned to *X, Y, k1, k2, k3* and CNP is necessary. In the actual tests a range of values starting with (k1, k2, k3) = (1,1,1) were experimented with. The results indicated a very high score for a number of common NPs, which were not semantically rich in terms of their ability to represent the 'aboutness' of the text. The value of k2 was gradually increased until some of the very common NPs were eliminated from the output. Once this was achieved, k3 was gradually increased until the output showed good NPs. High CNP values for NPs at levels 1b, 2 and 3 (see table 2) were arrived at on the basis of actual examination of several texts in the corpus, which showed the occurrence of good and useful NPs in those positions. Results of experiments that were conducted with a corpus of documents in the domain of Information science two sets of values for these are presented here. The two sets of values employed in the experiments conducted are shown in Table 1.

Constants	*Description*	*Set of values in the first experiment*	*Set of values in the second experiment*
X	X is the maximum number of occurrences that will be counted for a NP in a given text. Even if a NP appears more than X times, it is counted as X (to correct distortions)	10	7
Y	Minimum acceptable frequency of occurrence of a NP in a document to compute the number of documents in the corpus in which the NP occurs with a frequency >Y (for computing IDf (Y)	3	3
k1	Weight based on the frequency of NP in the document	1	1
k2	Weight (negative) based on the frequency of NP in the corpus of documents	10	15
k3	Weight based on the structure of the NP	10	15

Table 1. Parameter Values Employed

Category	Structure and Level of NPs	CNP value
1a	Level 1, structure (D*+ N)	0.25
1b	Level 1, any structure except (D* + N)	0.75
2	Level 2, any structure	1.0
3	Level 3, any structure	0.75
4	Level 4, any structure	0.5
>4	Level 5 or higher, any structure	0.25

(*D is any *determinant* such as '*a*', '*an*', '*the*', '*some*', '*few*', '*many*', etc)

Table 2. Assigned CNP Values

In order to understand how this categorization has been done, a few examples are presented below:

Category	Example of Text
CNP 1a:	A Informação (The information)
CNP 1b:	A informação correta (Correct information)
CNP 2:	O fluxo de informação (The flow of information)
CNP 3:	Estudos sobre o fluxo de informação (Studies on the flow of information)
CNP 4:	Autores dos estudos sobre o fluxo de informação (The authors of studies on the flow of information)
CNP 5:	Consensos entre os autores dos estudos sobre o fluxo de informação (Consensus among the authors of studies on the flow of information)

Table 3. NP Categories

Work is in progress on developing a new method of arriving at Score (NP) in which some of the arbitrariness of the present methodology is overcome. The new methodology (which will be reported shortly) proposes to employ no constants and use only actual data computed from the text and other texts in the corpus.

5. The Experiment

The experiments were conducted primarily to test the adequacy and utility of the methodology described in the foregoing sections. The corpora of texts used in the experiment consisted of sixty *e*-documents falling in the domain of Information science – all papers in two Portuguese language periodicals in the area of Information Science:

– Of the first 30 documents, 29 papers were from the journal *DataGramaZero*, and one paper from *Ciência da Informação*;

– The remaining 30 papers were slightly longer papers, all from the journal *Ciência da Informação.*

The decision to group the test documents in the corpora into two different groups was made with a view to examine the effects, if any, of the size of the document on the output.

The implementation of the methodology proposed here required a certain amount of computational work and also utilization of appropriate tools. The Figure 3 indicates the software tools utilized, the processes and stages involved as also the relationships between the processes.

Two important software tools that were utilized in the experiments are: 'PALAVRAS', a parser developed at the Southern University of Denmark, and 'PALAVRAS XTRACTOR', developed jointly by the Universidade do Vale do Rio dos Sinos (Unisinos) in São Leopoldo, Brazil, and Universidade de Évora, Portugal. The syntactically marked documents were presented as XML files. The tagged XML files were processed using a style sheet by the XML SPY software to create HTML files containing the extracted NPs. All the extracted NPs with a frequency higher than a pre-determined level (2 in the experiments) were processed using Microsoft Excel to compute their score and rank them. This was done using the formula explained in the section on methodology and the values defined for k1, k2 and k3.

5.1. Analysis of the Results

Table 4 presents an over of the output in terms of the number of NPs identified, the number of unique NPs and the number of NPs finally selected as descriptors based on the procedure developed in this research.

The score (NP) for every unique NP for every document was computed and the NPs for a document were ranked on this basis. For every document in the corpus about 1% of the NPs extracted from it were chosen from the ranked list of NPs (subject to a maximum of 15 NPs per document fixed purely as a convenience measure) for further analysis. In case two or more NPs obtained the same score (NP) and

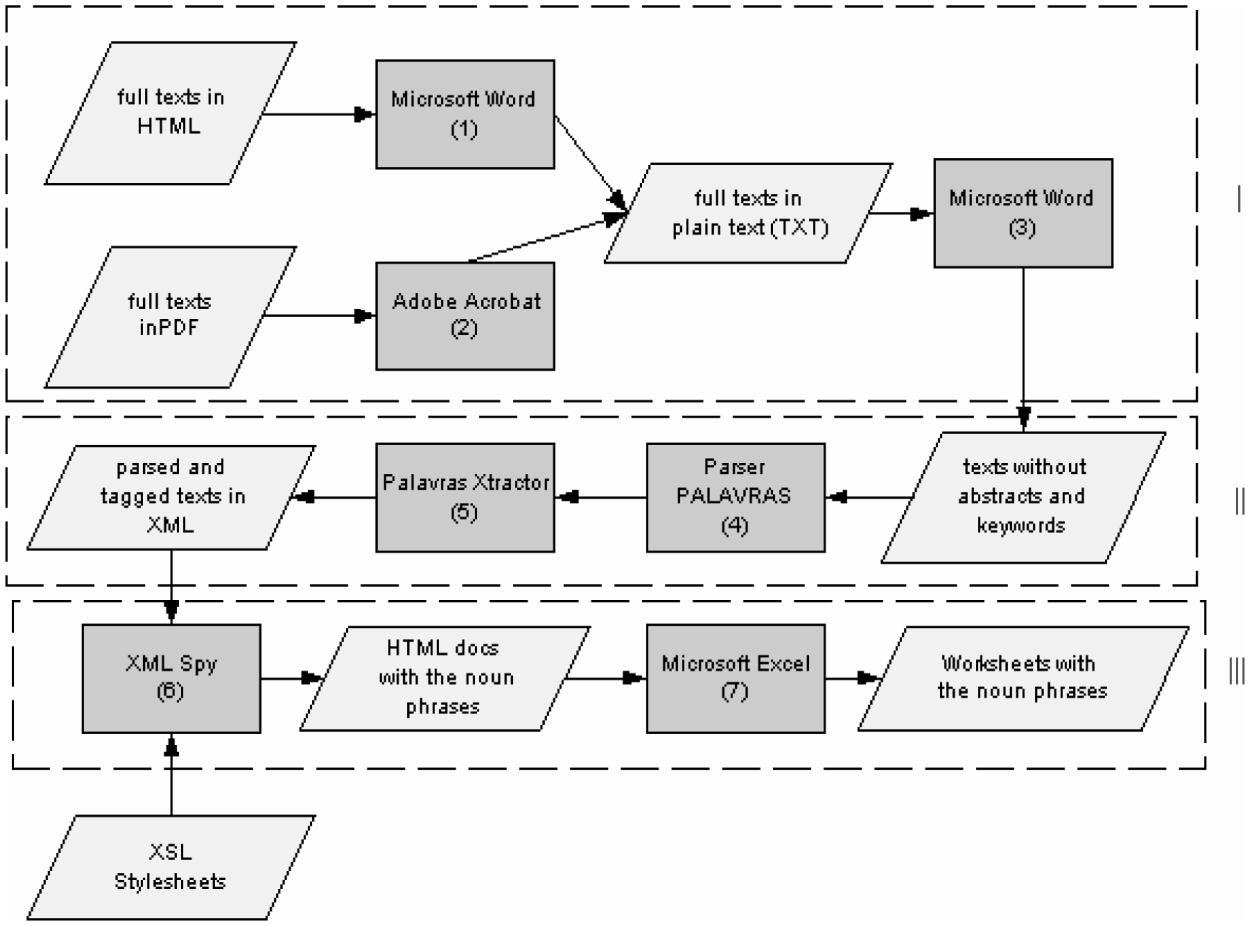

Figure 3. The Tools & Processes

one had to be selected as the descriptor, the issue was resolved as explained below:

- A NP that also formed the index term vocabulary of a thesaurus was preferred to one that was not;
- In case this did not help in resolving the issue the following criteria in that order was used:
  - NP with a higher frequency of occurrence in the document;
  - NP that was less evenly distributed among the documents in the *corpus*;
  - NP that belonged to a higher category based on its level and structure;
  - NP with more number of characters.

Doc #	Number of NPs			Doc #	Number of NPs		
	Identified	Unique	Selected		Identified	Unique	Selected
1	1673	1343	13	31	1702	1528	15
2	842	711	8	32	1902	1213	12
3	783	680	8	33	1941	1290	13
4	801	688	8	34	1480	1231	12
5	1478	1252	13	35	1011	788	8
6	984	836	8	36	735	552	8
7	638	521	8	37	2054	1382	14
8	779	684	8	38	772	624	8
9	1104	932	9	39	1873	1284	13
10	1146	1035	10	40	1156	962	10
11	619	554	8	41	1008	792	8
12	791	626	8	42	1244	1002	10
13	1342	1113	11	43	1808	1325	13
14	923	747	8	44	1375	1145	11
15	1063	877	9	45	1420	1176	12
16	888	810	8	46	1829	1453	15
17	1201	1084	11	47	987	810	8
18	5686	4287	15	48	1498	1223	12
19	1094	899	9	49	884	760	8
20	1299	1039	10	50	852	677	8
21	733	616	8	51	1225	1009	10
22	1837	1368	14	52	547	483	8
23	796	699	8	53	1364	1062	11
24	2048	1434	14	54	1535	1174	12
25	1368	988	10	55	1144	840	8
26	1246	1058	11	56	1386	1119	11
27	1173	971	10	57	1702	1353	14
28	788	667	8	58	1497	1166	12
29	617	539	8	59	733	632	8
30*	633	506	8	60	1702	951	10
%age	100%	81.28%	0.98%	%age	100%	76.81%	1.03%

Table 4. Summary of Output

NPs Extracted for							
		Texts 1 to 30 from *corpus*			Texts 31 to 60 from *corpus*		
I	First Experiment	NPs***	138	47.75%	NPs***	179	55.59%
		NPs**	66	22.84%	NPs**	63	19.57%
		NPs*	58	20.07%	NPs*	58	18.01%
		NPs–	27	9.34%	NPs–	22	6.83%
		SW	19	6.17%	SW	17	5.01%
II	Second Experiment	NPs***	137	47.40%	NPs***	173	52.58%
		NPs**	64	22.15%	NPs**	64	19.45%
		NPs*	56	19.38%	NPs*	64	19.45%
		NPs–	32	11.07%	NPs–	28	8.51%
		SW	5	1.70%	SW	7	2.08%

Legend: *NPs*** = Highly relevant NPs; NPs** = Reasonably relevant NPs*
NPs* = Moderately relevant NPs; NPs– = Non-Relevant NPs; SW = Stop words

Table 5. Output Categorized

The chosen NPs were manually examined for their relevance in terms of appropriateness and suitability for use as descriptors for the concerned document, and were categorized in terms of their degree of appropriateness. Table 5 presents an overview (summary) of the results.

As can be seen from the table, the results were quite satisfactory for the corpus of texts tested in the experiments. If both highly relevant and reasonably relevant NPs are considered, about 70% of the NPs extracted were quite appropriate and could be considered good quality descriptors for the concerned documents. The summary of results of the first experiment carried out for the documents 31-60 in the corpus (longer texts) is graphically presented below.

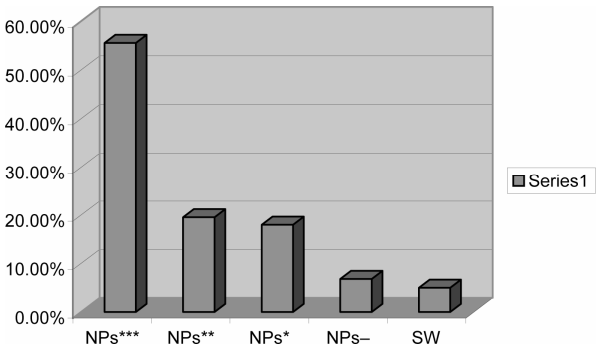

Figure 4. The Results

The results appear to suggest that the methodology does result in extracting NPs of value and utility for use as descriptors. The NPs that were arrived at for a text were compared with the keywords assigned by the authors of the text. The extracted NPs could be considered as having a higher information density than the keywords suggesting the utility of the methodology. Some conclusions could be drawn on the basis of the analysis of the results:

- The NPs retained the context in which a word occurred to a large extent. For example, the NP 'Rio de Janeiro' which is the name of a city would be extracted rather than 'Rio' (meaning a river) and 'Janeiro' (January), which may be totally irrelevant in the given context;
- As the NPs are not subjected to any stemming, it is possible to differentiate between some lexemes, e.g. 'gestao' (management) and 'gestor' (manager);
- At higher frequencies the qualitative advantages of the extracted NPs were quite visible. It does appear that in any frequency-based approach, NPs will be far more capable of representing the 'aboutness' of a document than high frequency keywords (e.g. "interface de consulta" rather than "interface" and "consulta").

These are strong reasons to seriously consider and further explore the methodology for possible refinements.

When the experiments were designed and it was decided to use two sets of texts (one set having longer texts), the idea was to see if there was any difference. It was thought that there is a strong possibility that identification of and discrimination between good NPs and not so good NPs would be easier in large texts. It can be seen from the data that while the average number of NPs extracted for the first 30 texts in the corpus is 1212, the corresponding figure for the texts 31- 60 in the corpus is about 1345, which is roughly 10% higher. The difference would have been even higher, had it not been for document #18 in the first set which happened to be a very long paper. This however, needs to be tested with a larger corpus and based on texts with substantial difference in their lengths.

6. Conclusions And Future Work

This research emerged largely from the realization of the near impossibility of manually organizing large collections of digital resources. The central objective of the work was to propose an effective mechanism for extracting semantically rich NPs that could serve as descriptors to represent the 'aboutness' of documents from which they are extracted. The methodology employed which involves, frequency data for a NP within a text, data about the number of documents in the corpus that contain the NPs, and structure and level of the NPs appears to yield reasonably good results as shown above. The process of testing the methodology with a larger corpus and further refinement of the methodology, especially that related to computing the Score (NP) is in progress. The results available with us now appear to contradict the findings of declaredly unsuccessful previous experiences, which sought extraction of descriptors based on syntactic structures of the sentences (Earl 1970, Paice 1981, Fum et al. 1982, and Lancaster 1993, 250-51). Probably an important factor is the fact that tools for automatic extraction of NPs were not many and these have become more widely available only in the last one decade and more. Although Kuramoto (1999, 2003) reported a study on the utility of NPs in IR in Portuguese, we have not found any sign of follow-up of those studies in the Brazilian scientific literature. It is expected that the methodology presented here – and others that may derive from it – will be useful in situations where documents are added at a rate that makes manual processing extremely difficult. We are currently working along some of the paths opened by the methodology, and the outcome and refinements to the methodology will be reported when some more results become available. Work is also in progress with regard to building domain-specific, open and dynamic 'stop lists' consisting of extracted phrases that are not useful as descriptors. While this will require some manual intervention in the initial stages, it is expected that over a period of time after processing a reasonable number of documents in a domain, the 'stop list' will grow to a level to be able to handle most of such NPs without human intervention. It is also possible to build into the system a validation process based on authorities. For example, it can be argued that a NP that is also part of the vocabulary of a standard thesaurus in the domain is likely to be a useful descriptor and based on this a validation process could be built into the methodology. Enhancements to the parser and other possible applications of the output are also being explored.

References

Baeza-Yates, Ricardo and Ribeiro, Berthier de Araújo Neto. 1999. Modern information retrieval. New York: ACM Press.

Bick, Eckhard. Parsers and their applications. http://www.hum.au.dk/lingvist/lineb/home_uk.htm.

Bick, Eckhard. 1996. Automatic parsing of Portuguese. In García, Laura Sánchez, ed., *Anais / II Encontro para o Processamento computacional de português escrito e falado*. Curitiba: CEFET-PR. http://beta.visl.sdu.dk/~eckhard/pdf/curitiba.ps.pdf.

Bick, Eckhard. 2001. The VISL System: research and applicative aspects of IT-based learning. The VISL System: Research and applicative aspects of IT-based learning. In: *Proceedings of NoDaLiDa 2001 (Uppsala)*. http://beta.visl.sdu.dk/~eckhard/pdf/NoDaLiDa2001.ps.pdf.

Gasperin, Caroline Varaschin et al. 2003. Uma ferramenta para resolução automática de correferência. In Anais do XXIII Congresso da Sociedade Brasileira de Computação, VI Encontro Nacional de Inteligência Artificial, Vol VII. Campinas.

Gasperin, Caroline Varaschin et al. 2003. Extracting XML syntactic chunks from Portuguese corpora. In Proceedings of the Workshop on Traitement automatique des langues minoritaires 2003 Natural Language Processing of Minority Languages and Small Languages - Batz-sur-Mer France June 11 – 14, (2003).

Projeto DIRPI. 2001. Desenvolvimento e integração de recursos para pesquisa de informação. Cooperação Científica e Técnica Luso-Brasileira. ICCTI/ GRICES-CAPES, Universidade de Évora, Universidade Nova de Lisboa, Unisinos, PUC-RS.

Kuramoto, Hélio. 1995. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. Ciência da Informação, Brasília, v. 25, n. 2. Artigos. Also available: http://eprints.rclis.org/archive/00003571/01/ Kuramoto1995.pdf.

Kuramoto, Hélio (1999). Proposition d'un système de recherche d'information assistée par ordinateur avec application à la langue portugaise. Tese (Doutorado em Ciências da Informação e da Comunicação) – Université Lumière - Lyon 2, Paris, França.

Lancaster, F. W. 1997. Indexação e resumos: teoria e prática. Brasília, Briquet de Lemos, 1993.

Liberato, Yara G. 1997. A estrutura do sintagma nominal em Português: uma abordagem cognitiva.. 203 f. Tese (Doutorado em Letras) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte.

Miorelli, S. T. 2001. Extração do sintagma nominal em sentenças em Português.. 98 f. Dissertação (Mestrado em Ciência da Computação) – Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre.

Moreiro, José et al. 2003. Desarrollo de un método para la creación de mapas conceptuales. Belo Horizonte : Anais do ENANCIB.

Perini, Mário A. 1985. A gramática Gerativa: introdução ao estudo da sintaxe portuguesa. 2a edição. Belo Horizonte: Vigília.

Perini, Mário A. 1995. Gramática descritiva do português. 2a edição. São Paulo: Editora Ática.

Perini, Mário A. et al. 1996. O SN em português: A hipótese mórfica. Revista de Estudos de Linguagem – UFMG, Belo Horizonte Julho/Dezembro: 43-56.

Rossi, Daniela, et al. 2001. Resolução automática de Correferência em textos da língua Portuguesa. REIC Revista de Iniciação Científica da SBC, http://www.sbc.org.br/reic/, v. 1, n. 2.

Ruwet, Nicolas. 1975. Introdução à gramática gerativa. São Paulo: Perspectiva, Editora da Universidade de São Paulo.

Salton, Gerard & Mcgill, Michael J. 1983. Introduction to modern information retrieval. New York: McGraw-Hill.

Sant'anna, Victor, et.al. 2000. Cálculo de referências anafóricas pronominais demonstrativas na língua portuguesa escrita. 100 f. Dissertação (Mestrado em Informática) – Instituto de Informática da PUC-RS – Porto Alegre.

Souza, Renato Rocha. 2005. Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais.. 214 f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência de Informação, Universidade Federal de Minas Gerais, Belo Horizonte.

Sparck Jones, K. & Willett, P. eds. 1997. *Readings in information retrieval*. San Francisco: Morgan Kaufmann.

Velumani, G. & Raghavan, K.S. 2005 Automatic extraction of keywords from Web resources. *Information studies* 11(3): 185-94.

Velumani, G. & Raghavan, K.S. 2006. Extraction of keywords: a noun phrase-based methodology. In Knowledge representation and information retrieval ed. K. S. Raghavan. Bangalore: DRTC, Indian Statistical Institute, paper P.

Vieira, Renata. 1998. A review of the linguistic literature on definite descriptions. *Acta semiotica et lingvistica* 7: 219-58.

Vieira, Renata. et al. 2000. Extração de sintagmas nominais para o processamento de co-referência.. Anais do V Encontro para o processamento computacional da Língua Portuguesa escrita e falada PROPOR, 19-22 Novembro Atibaia SP.

VISL. About VISL. available at http:// visl.hum.sdu.dk/visl/about/index.html.

Woods, John O, Richardson, Ryan & Fox, Edward A. 2005. Multilingual noun phrase extraction using a part-of-speech tagger; available at http://www. writing.eng.vt.edu/Abstract/John%20Woods.pdf