A methodology for automatic indexing using noun phrases

Renato Rocha Souza[1], K.S. Raghavan[2]

[1] 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

[2]Documentation Research & Training Centre, Indian Statistical Institute, Mysore Road, Bangalore 560 059, India.

rsouza@eci.ufmg.br, ksraghav@hotmail.com

Abstract. *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

Considering the sheer volume of information to be handled today, it is widely recognized that solutions to handle such a situation should necessarily be technology-based and make effective use of intelligent technologies. 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. This paper deals with a strategy to index automatically using Noun Phrases.

2. Noun Phrases and Information Retrieval Systems

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, pp.169-170). 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.

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, p.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), 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 "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". (1996, p.6) 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.

3. Objectives

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 / phrase extracted from a text to accurately and meaningfully represent full or partial subject content of the text. This paper reports a research that seeks to carry forward utilization of semantics embedded in texts in Portuguese language for deriving meaningful representations of their 'aboutness'. It explores 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 hypotheses that 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, and 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 1 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 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 frequency 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. 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:

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

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.

Table 1. Parameter Values Employed

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 2. Assigned CNP Values

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)

In order to understand how this categorization has been done, a few examples are presented below in Table 3.

Table 3. NP Categories

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)

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 1 indicates the software tools utilized, the processes and stages involved as also the relationships between the processes.

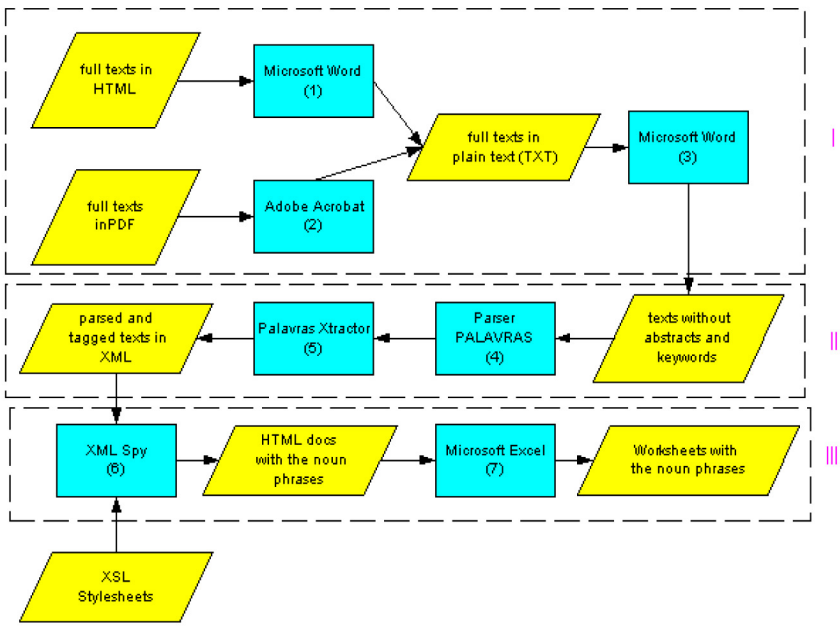

Figure 1. The Tools & Processes

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

The table below 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.

Table 4. Summary of Output

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%

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 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.

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. The following table presents an overview (summary) of the results.

Table 5. Output Categorized

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

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.

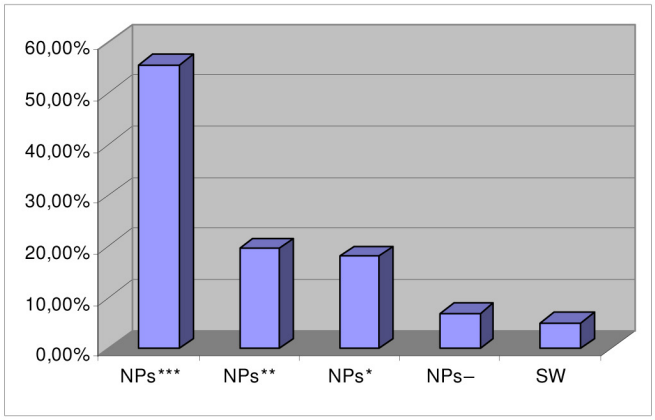

Figure 2. The Results

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, pp. 250-251). 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, R. & Ribeiro-Neto, B. (1999). "Modern Information Retrieval". New York: ACM Press, 511p.

Gasperin, C.V. et al (2003). "Extracting XML chunks from Portuguese corpora". In: Proceedings of the Workshop on Traitement automatique des langues minoritaires.. Batz-sur-Mer.

Kuramoto, H. (1996). “Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais”. Ciência da Informação, Brasília, v. 25, n. 2. Also available: http://www.ibict.br/cionline/250296/25029605.pdf

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

Sparck Jones, K. & Willett, P. Ed. (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), 185-194

Velumani, G. & Raghavan, K. S. (2006). “Extraction of Keywords: A Noun Phrase-based Methodology” (In Knowledge Representation and Information Retrieval edited by K. S. Raghavan. – Bangalore: DRTC, Indian Statistical Institute, paper P.)

Woods, J.O.R.R. & Fox, E.A. (2005). “Multilingual Noun Phrase Extraction Using a Part-of-Speech Tagger”. Available at http://www.writing.eng.vt.edu/Abstract/John%20Woods.pdf