Using Noun Phrases for Local Analysis in Automatic Query Expansion

João Marcelo Azevedo Arcoverde
Maria das Graças Volpe Nunes
Departamento de Ciências de Computação e Estatística
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo - Campus de São Carlos
Caixa Postal 668, 13560-970 - São Carlos, SP - Brasil
{jmaa,gracan}@icmc.usp.br

Wendel Scardua
Departamento de Ciências de Computação e Estatística
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo - Campus de São Paulo
Caixa Postal 66.281, 13083-970 - São Paulo, SP - Brasil
articuno@ime.usp.br

Abstract

Blind relevance feedback constitutes a widely used technique to improve the performance of IR systems. Its result is directly influenced by the correct choice of the potentially qualified features used to expand the initial query. It is well known the power of noun phrases in the role of descriptors with high discriminatory and informative potential. This work presents a local analysis technique to automatic query expansion through the use of noun phrases extracted from the pseudo-relevant set, for the Ad-Hoc, monolingual (Portuguese) track. Even though the technique is language independent, specific resources for Portuguese were used to the noun phrases extraction through the use of Machine Learning techniques. In our experiments with a specific IR system, it has been observed an improvement in the results obtained over 38% of the topics, and also its depreciation over some others topics. However, this fact constitutes a clear evidence of how the use of NLP techniques can influence such processes, showing real possibilities for improving the presented technique.

Keywords

Noun phrases, blind relevance feedback, local analysis, query expansion

1 Introduction

Natural language is characterized by the imprecision of the terms used in the text, which constitutes the main vehicle of acquisition and dissemination of the human knowledge. Interpretation and meaning attribution are almost always defective and ambiguous processes. The information retrieval systems available nowadays reflect the linguistic phenomenon of this nature, especially related to the user's expectation of coherent answers to his/her information needs. The query formulation process constitutes a challenge to these systems.

In a IR system, the query is defined as the elaboration process of the user information need. For the IR models most commonly used in Web environment, the query formulation through the use of keywords appears as the main language for human-machine communication process. It is a intuitive language, of easy manipulation and allows the ordering of the set of documents returned by the query according some relevance judgment. This ordering is a difficult task, either because of the user's inability to efficiently articulate his/her information need, or the own nature of the human language.

Besides the difficulty of query elaboration, the process to select relevant documents normally involves interactive cycles between the user and the system, including, most of times, reformularizations of the initial query. One strategy to simplify this process is to expand the initial query with related terms, trying to feed the system with a more elaborated context, minimizing the problems due to the human language.

The remaining of this article is as follows: Section 2 presents a broad vision of the query expansion problem such that we can put our technique among the most known approaches; Section 3 presents a method for noun phrase extraction based on machine learning approach; Section 4 describes our experiment with pseudo-relevance feedback; Section 5 presents the evaluation of the method according well established metrics for IR systems; and Section 6 concludes the article with some observations.

2 Query Expansion

The process of reformulating the initial query must address a selection criterion to choose the new descriptors that will compose the expanded query, as well as a strategy to recalculate its weights together with those from the initial query. The decision of how many descriptors to pick up is a problem one must analyze experimentally. The point is that the descriptors must convey a qualitative semantic power that distinguish them from the others, towards within the context where they were identified.

The query expansion can be done through the use of the entire collection of documents or through an external knowledge database. Although many researchers succeeded in the use of external databases to query expansion [4, 7], the cost of its attainment and maintenance generally restrict them to domain specific applications.

There are basically two main approaches to query expansion: a) interactive - when the user interacts with the system, feeding information about the relevance of the returned documents; and b) automated - when there is no interaction from the user throughout the process. In the first approach, one say that there is relevance feedback. The second approach (automated), it has been said there be pseudo-relevance feedback. The scope of analyzed documents used to expand the initial query can be global, when the entire collection is garned, or local, when only a subset of the collection is used, mainly the top "n" ordered documents returned from the initial query.

The automated and local analysis constitutes a tendency to automate the query expansion process for domain independent collections of reasonable dimension, producing improvements to the recall and precision of IR systems [10]. It has the advantage of the exploratory power of the local context supplied by the query, becoming more appropriate than the global analysis.

3 Noun Phrases identification

The noun phrases are broadly known as the set of elements that referrers to concepts, objects or facts from the real world and, therefore, carry high discriminatory information [6]. These linguistic structures have been widely used in many computational problems, for example, in a controlled vocabulary indexing procedure. Here they are used to the query expansion process in a specific IR system, representing the most important descriptors from which the new expanded query will be constructed.

The problem of recognizing and extracting the noun phrases of free texts have been evolved from the use of symbolic computational programs to the statistical ones. In part the reason from which this evolution can be explained is attributed to the matureness of the supervised Machine Learning (ML) techniques, as soon as trustworthy examples are presented.

These ML techniques have surpassed the performance of the applications that employs manually created and maintained grammatical rules, what has been characterized as a hassle process and, in many situations, do not easily share between different applications.

This work considers only the lexical noun phrases - those which the nucleus is a name. We had used the system developed by [9], based on TBL (*Transformation Based Learning*) [1], customized for the Portuguese language.

The idea behind the algorithm is to generate an ordered list of transformation rules, that gradually fixes errors in a training corpus, produced by an inexact initial classification. From each new iteration, the rule chosen to compose the list of learned rules is that which will trigger more error reduction in the training corpus classification, as shown in Figure 1.

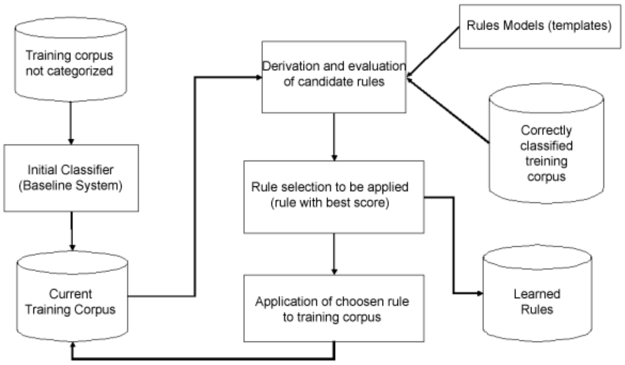
Training corpus not categorized
Initial Classifier (Baseline System)
Current Training Corpus
Derivation and evaluation of candidate rules
Rules Models (templates)
Correctly classified treining corpus
Rule selection to be applied (rule with best score)
Application of choosen rule to training corpus
Learned Rules

Figure 1: Transformation Based Learning

4 Experimental setting

Our team (NILC[1]) has participated in Clef 2006 for the Ad-Hoc, monolingual, Portuguese language track for the very first time, with its IR system designed to explore the power of noun phrases as the main descriptors of an hybrid indexing scheme, which aims to combine statistical and linguistic knowledge extracted from the corpus.

The method presented in our two runs submitted to the task explored both how the hibrid text representation can produce good effectiveness and how query expansion could be done within this model. This report emphasizes the second point, for two reasons: *i*) query expansion over a linguistically motivated index is a completely new branch of research to our actual context; and *ii*) the two runs submitted to the track can be compared each other to evaluate the performance the second run took over the first one, which has not used query expansion. Both runs used the same indexing scheme.

For the query expansion we used pseudo-relevance feedback to automatically expand the initial query without demanding interactions between the user and the whole process. The local analysis was done within top twenty first documents returned from the initial query, ordered by a slight variation of the Okapi BM25[8] ranking algorithm. This value is empiric and can vary according to the collection itself and to the topic and relevance judgments related to it.

[1] nilc.icmc.usp.br

4.1 Collection, topics and initial queries

In Clef 2006, two collections were disposed for the Ad-Hoc, monolingual for Portuguese track, including issues from Brazil and Portugal, that differ each other by many peculiarities that must be addressed in a linguistic context, varying from typography differences within terms, from other word mismatch phenomenon (synonym and polissemy). These variations can influence the initial query precision and recall, thus the overall query expansion process. The corpus is composed by four collections: Folha94, Folha95, Publico94 e Publico95[2], totalizing 210.736 free text documents (approximatelly 560 Mb).

Besides, it was disposed 50 search topics from different subjects, along with their respective descriptions. These descriptions were manually processed to derive the initial queries set, one for each topic. Each initial query is composed by one Boolean expression over the main descriptors identified along the description text.

One important operation named "term proximity" was implemented, once it is useful to store the absolute position for each document term at indexing time. This allow to compute the following Boolean expression: $C_n = +d_1 \ -d_2 \ + (d_3 \ \backslash n \ \ d_4)$, where $n \in Z$. In this query, C_n searches for all documents that holds the term d_1 and do not hold the term d_2 and hold d_3 and d_4 distant from each other, at most, n terms, no metter the order between them. This operation is essential to work with the noun phrases nucleus, considering some relative term distance between them, assuming that this distance is a clue to its contextual correlations.

4.2 Pre-processing the collection

The collection required three different levels of pre-processing before indexing, which took approximately 70 hours using 4 dedicated Pentium-IV machines (3.2 GHz CPU and 2Gb RAM), one for each collection. This was a very challenging step of building a linguistically motivated index.

First of all, the text was delimited by sentences, structuring one sentence per line. Following this segmentation, each term from each sentence were *tokenized* and morphologically analyzed. In this phase we proceed with the morphologic disjunctions, for example, "do = de + o", "àquele = a + aquele", "dentre = de + entre", etc.

In the second level of pre-processing, the text was *POS-tagged* using a language modeling proposed by *MXPOST*[3], associating each token to its grammatical function based on its context evidences. The training corpus used to induce the classifier holds 41.883 sentences extracted from Mac-Morpho[4], from the LacioWeb[5] project.

The third level of our pre-processing architecture was responsible to identify and tag the noun phrases for each labeled sentence, for each document of the entire collection. Therefore, it also flagged the nucleus for each noun phrase (which can be formed by multiple lexical words). It was used the *TBL* (Transformation Based Learning) algorithm [1], as we have already briefly described in the last Section. It was used a corpus of 4.393 sentences selected from *Mac-Morpho*[6], which were thoroughly revised by [2]. Following [9], the process described to identify the noun phrases reaches approximately 87% of *F-measure*.

4.3 Controlled indexing vocabulary

Once the collection is pre-processed, the indexing phase takes place. It requires basic linguistic operations such as case-folding, accents mapping and *stopwords* removal. Once the terms are syntactically labeled, it is possible to take some decisions in order to reduce the size of the space-terms through the use of a controlled vocabulary indexing scheme. For example, numerical sequences are not indexed, unless if they were part of some noun phrase nucleus. Analogously, we can treat

[2] complete editions from 1994 and 1995 of journals PÚBLICO (www.publico.pt) and Folha de São Paulo (www.folha.com.br), compiled by Linguateca (www.linguateca.pt)

[3] Maximum Entropy, de Adwait Ratnaparkhi

[4] http://www.nilc.icmc.usp.br/lacioweb/corpora.htm

[5] http://www.nilc.icmc.usp.br/lacioweb/

[6] http://www.nilc.icmc.usp.br/lacioweb/corpora.htm

the same way monetary values, percentages, etc. All the verbs were indexed in their infinitive form, that measure saved a large amount of disk space, second only to the numerical sequences. In our indexing scheme, all the multi-word noun phrases were also indexed as a single descriptor, together with their unigram terms. This approach takes an alternative fast ranking algorithm to weight these structures at search time.

4.4 Pseudo-relevance feedback

The proposal of the method is to free the user from interacting with the system, even though this interaction may happen spontaneously at the time the system shows the new reformulated query, before its re-submission through an interface. At this point the user visualizes the new Boolean expression, optionally change and re-submit to the system, giving turn to the expansion interaction. This submition could be completely automated and transparent to the user. However, for experimental reasons, we have decided to track the way the query was reformulated and have the power to influence it.

One relevant challenge is the exploration of alternatives to identify which noun phrases supplies the best context to efficiently contribute to query reformulation. One important problem is to choose which parts of the documents from which will be extracted the noun phrases that will act as potential candidates as descriptors in the expanded query. We can extract from the entire document or from the top relevant passages. We observed that the documents from the collection share a uniform size, as well each document reflects only one topic, with some exceptions. We have decided for a third alternative: to select only those candidates that are near to the signaled occurrence triggered from the initial query. This is because words with similar meanings tend to occur in similar contexts[5]. Thus, the objective is to minimize the noise generated from those descriptors that are far from the context and, therefore, probably refer to a different topic. Thus we have extracted all the noun phrases from the sentence where exists at least one match triggered from the Boolean search expression, as well from the adjacent sentences (one above and one below). These values are also parameterized into the system and can vary among the experiments.

In order to select an enough amount of descriptors (also determined experimentally) to compose the new query, we attributed to the noun phrases weights that reflect their evidence in the text. To calculate the weight of the noun phrase, only the nucleus was considered, discarding their determinants and modifiers.

The weight of a noun phrase s in a document d follows the Equation (1), according to [3].

$$w_{s,d} = f_{s,d} \times \sum_{i=1}^{n} w_{t_i,d} \tag{1}$$

where:

- $f_{s,d}$ is the frequency of occurrence of s in d and;
- $w_{t_i,d}$ is the weight of the n-th term t_i from the s nucleus in d;

Each noun phrase chosen from the sentences has its nucleus splited by unigram terms. These terms suffer a lemmatization process to provide a natural conflation among them. The lemas extracted from the nucleus produces a better weight schema than if it were done without lemmatizing the terms.

The lemmas are weighted according to their frequency in the noun phrases of the document. Optionally, we can multiply this value by a factor *idf*, that measures the rarity of this lemma in the pseudo-relevant set. The frequency of occurrence of the noun phrase s in d is the sum of how many times this multi-term structure occurs in the document.

Once the weight of each lemma had been calculated, as well the frequency of each noun phrase s in document d, the weight of s in d is the product of its respective frequency by the sum of the weight of their lemmas. Then it is possible to rank the set of noun phrases from the pseudo-relevant documents according to its weight, in descendent order, and pick up the first top n noun phrases. These are the descriptors that will compose the new expanded query, rearranged in a new Boolean search.

5 Evaluation

Each query (expanded or not) formulated over one topic returns a set of documents ordered by some relevance criterion. Each returned document is a record that obeys a predefined output layout to be submitted to a specific system that will judge the correctness of the claimed relevance. The set of all records grouped by topic constitutes a run. Each run reflects the behavior of the IR system for all disposable topics.

The presenting work generated two different runs to be evaluated against the relevance judgments by the Clef team. Also, the runs shall be compared one against another, which constitutes the objective of this article. They are: *i*) NILC01 - all the initial queries (one for each topic) with no expansion, and *ii*) NILC02 - with query expansion. The runs evaluated by the Clef judges reported metric values that match those evaluated by the *trec_eval*[7] program, using the same relevance judgments. This is useful for future manipulations on the process by our team, without depending on external procedures, until the next Clef campaign.

We have focused on the traditional metrics used to evaluate IR system: *i*) *MAP* (*Mean Average Precision*) - that express the mean of the precision after each relevant document has been retrieved. This metric emphasizes the earlier relevant documents retrieved; *ii*) precision - expresses how many relevant documents were retrieved in relation to the number of retrieved documents; *iii*) recall - expresses how many relevant documents were retrieved in relation to the entire collection.

Only 19 from 50 topics (38%) have expressed better MAP compared to the first run (initial queries). There was only one draw for one topic that did not return results in the first run and, therefore, could not be expanded. It was verified that 30 topics from the second run presented a loss of precision (MAP) compared to the first run. This means that, despite of expansion had presented more relevant documents for the majority of topics, it also returned much more irrelevant documents over time, scattering the relevant ones among them, harming the ranking for the returned set. This justifies the loss of precision at interpolated levels of recall.

The MAP metric for both runs can be expressed by topic in a bar chart, as shown in Figure (2). The NILC01 MAP is of 35.20%, and for the NILC02 run is of 29.01%. The precision and recall are mapped in an area chart, as shown in Figure (3), that figures out the trade-off between precision for each level of recall, in a percentage scale, for all topics.

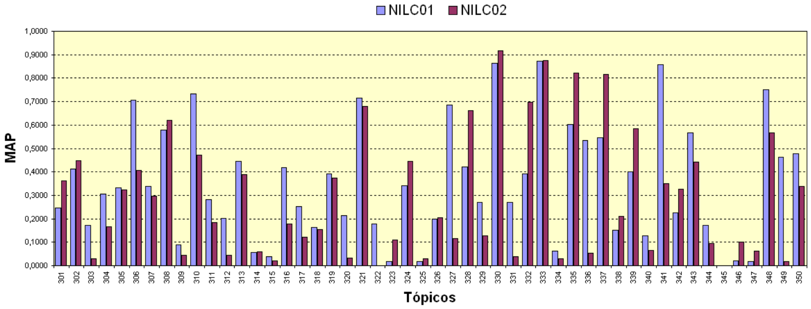

Figure 2: MAP over topics

No intervention was made in the parameters that prevails the behavior of the IR system while NILC02 run was processing all topics at once. After the experiment was submitted, it was perceived that the initial query quality is the main factor to influence the query expansion. There are others factors that intervenes in the process over each topic, such as *i*) the number of noun

[7] Chris Buckley - http://trec.nist.gov/

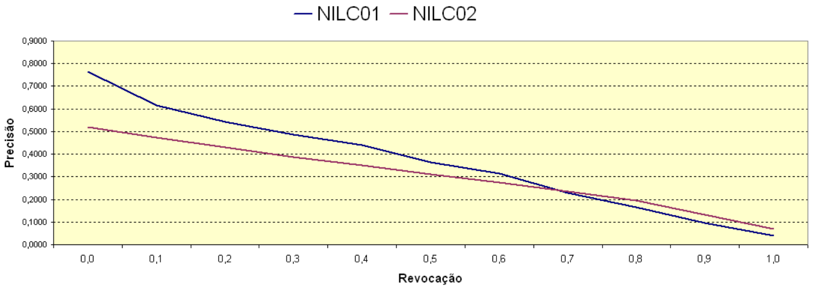

Figure 3: Precision at each 10% of recall

phrases chosen; *ii*) the number of chosen sentences to extract the noun phrases and *iii*) the number of documents which delimits the pseudo-relevance set.

6 Conclusion

The IR field has always used basic NLP techniques to aid document structuring process. However, only in the past few years researches have pointed out advances related to a more sophisticated generation of NLP techniques that justify the cost for its use, comparing to the traditional approaches.

This work investigated evidences that serves as a base to the hypothesis that applying linguistic knowledge methods is viable, contributing to the traditional statistical methods available. It was presented one technique of local analysis for query expansion without user intervention, according to a linguistically motivated model based on noun phrases. These structures carry information with a highly discriminative power, therfore playing a better role as descriptors in text representation models.

The obtained results encourage us to the individual manipulation of each expanded query for each topic before submitting it to the system. This may allow achieving better combinations of system parameters, revealing more conclusive results regarding the experiment.

The high computational cost (time and space complexity) demanded by the preprocessing and indexing stages allow the use of linguistic resources on appropriate data structures to be better explored by the user at search time. The time for expanding the query triggered at execution phase, using previously indexed linguistic knowledge, is highly acceptable and does not negatively intervene in user experience.

There are open research possibilities to explore how other processes could be benefited by the use of linguistically motivated text representations, using noun phrases, for example, specially for the Portuguese language. One example could be to evaluate the impact of these structures in automatic text categorization processes, that can be used to filter irrelevant documents at search time, contributing to increase the effectiveness of such IR systems.

References

[1] E. Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565, 1995.

[2] M. C. Freitas, M. Garrãoo, Oliveira C., C. N. Santos, and M. Silveira. A anotação de um corpus para o aprendizado supervisionado de um modelo de sn. In *Proceedings of the III TIL / XXV Congresso da SBC*, São Leopoldo - RS, 2005.

[3] M. Gonzalez. *Termos e Relacionamentos em Evidência na Recuperação de Informação*. PhD thesis, Universidade Federal do Rio Grande do Sul (UFRGS), 2005. Tese de Doutorado.

[4] M. A. I. Gonzalez and V. L. Strube de Lima. Recuperação de informação e expansão automática de consulta com thesaurus. In *XXVII Conferência Latinoamericana de Informática (CLEI'2001)*, pages 1–10, Mérida, Venezuela, 2001.

[5] Z. Harris. *Mathematical Structures of Language.* New York - USA, 1968.

[6] H. Kuramoto. Sintagmas nominais: uma nova proposta para a recuperação de informação. *DataGramaZero - Revista de Ciência da Informação*, 3(1), 2002.

[7] L. A. S. Pizzato and V. L. Strube de Lima. Evaluation of a thesaurus-based query expansion technique. In N. J. Mamede, J. Baptista, I. Trancoso, and M.G. Volpe Nunes, editors, *Proceedings of the 6th Workshop on Computacional Processing of the Portuguese Language - Written and Spoken. Lecture Notes in Computer Science 2721*, pages 251–258, Universidade do Algarve-FCHS, Faro, Portugal., June 2003. Springer-Verlag.

[8] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR*, pages 232–241, 1994.

[9] C. N. Santos. *Aprendizado de Máquina na identificação de sintagmas nominais: o caso do português brasileiro.* PhD thesis, Instituto Militar de Engenharia (IME), Rio de Janeiro, 2005. Dissertação de Mestrado.

[10] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, 2000.