The PUCRS-PLN Group participation at CLEF 2006

Marco Gonzalez and Vera L. S de Lima
Grupo PLN – Faculdade de Informática – PUCRS
Av. Ipiranga, 6681 – Prédio 16 - PPGCC
90619-900 Porto Alegre, Brazil
{gonzalez, vera} @inf.pucrs.br

Abstract

This paper presents the 2006 participation of the PUCRS-PLN Group in CLEF Monolingual Ad Hoc Task for Portuguese. We participated with the TR+ Model based on nominalization, binary lexical relations (BLR), Boolean queries, and the evidence concept. Our alternative strategy for lexical normalization, the nominalization, is the transformation of a word (adjective, verb, or adverb) into a semantically corresponding noun. BLRs, which identify relationships between nominalized terms, capture phrasal cohesion mechanisms, like those that occur between subject and predicate, subject and object (direct or indirect), noun and adjective or verb and adverb. In our strategy, an index unit may be a single term or a BLR, and we adopt the evidence concept, i.e., the index unit weighting depends on the occurrence of phrasal cohesion mechanisms, besides the frequency of occurrence. We detail here these features, which implement lexical normalization and term dependence in an information retrieval system.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H3.3 Information Search and Retrieval

General Terms

Search engine, information retrieval evaluation, lexical normalization, term dependence

Keywords

CLEF, ad hoc

1 Introduction

The PUCRS-PLN Group participated in CLEF 2006 in Monolingual Ad Hoc Portuguese Retrieval Task with the trevd06 run using manual query construction from topics. This run adopts the TR+ Model [5].

TR+ Model is based on nominalization [6, 7], binary lexical relations (BLRs) [4, 6], Boolean queries [5], and the evidence concept [4]. Nominalization is an alternative strategy used for lexical normalization. BLRs, which identify relationships between nominalized terms, and Boolean queries are strategies to specify term dependences. The evidence concept is part of TR+ Model for term weighting using word frequency and phrasal cohesion mechanisms [8]. Trevd06 run uses a probabilistic approach for information retrieval.

In our strategy, a index unit (henceforth descriptor) may be a single term (e.g.: "house") or a relationship between terms (e.g.: "house of stone"). BLRs represent those relationships (e.g.: "of(house,stone)"). To each descriptor (a term or a BLR) is assigned a weight, an evidence in TR+ Model. Its evidence shows the importance of the concept that the term or the BLR describes in the text. Descriptors and their weights constitute the descriptor space.

This paper is organized as follows. Section 2 introduces the TR+ Model based on the nominalization process, the binary lexical relation recognition, a new term weighting schema based on the evidence concept, and the Boolean query formulation. Section 3 describes the collection, features of indexing files, and difficulties found. Section 4 shows the results of trevd06, and Section 5 presents final considerations.

2 TR+ Model overview

Figure 1 shows an overview of TR+ Model, including steps for descriptor space generation, in indexing phase, and relevance classification of documents, in searching phase.

In TR+ Model [5], documents and queries in natural language receive the same treatment in order to construct the descriptor space, in indexing phase, and to start the Boolean query formulation, in searching phase. First, in preprocessing step there are tokenization (words and punctuations are identified) and morphological tagging (morphological tags are assigned to each word or punctuation). Then, the nominalization process is performed for generating nominalized terms and, in the next step, BLRs are extracted.

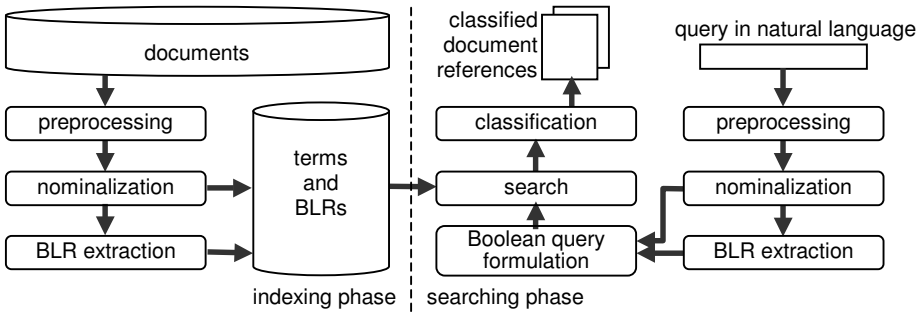

Figure 1. TR+ Model overview

In searching phase, we look for nominalized terms and BLRs recognized in the query, in the descriptor space. The document relevance values are computed according to descriptor weights (evidences) and to predefined Boolean operations included in the query. Finally, the documents are classified.

3.1 Nominalization process

Nominalization [7] is an alternative strategy used for lexical normalization. It is based on the fact that nouns are usually the most representative words of the document content [11], and because queries are usually formulated through noun phrases [9]. In our work, nominalization is understood as the transformation of a word (adjective, verb, or adverb) found in the text, into a semantically corresponding noun which appears in the lexicon.

Nominalization operations, according to TR+ Model, derive abstract and concrete nouns. Abstract nouns refer to events (e.g., "to meet → meeting"), qualities (e.g., "good → goodness"), states (e.g., "free → freedom"), or other abstract entities, which can be derived from adjectives, participles, verbs, or adverbs. Concrete nouns, on the other hand, refer to agents mostly derived from verbs (e.g., "to build → builder"), or something that is involved or associated with an entity, mainly derived from adjectives (e.g., "numerical → number").

To develop this idea, an automatic nominalization process was implemented and integrated to our indexing strategy. We developed the tools FORMA and CHAMA that automatically derive nominalized terms from a Brazilian Portuguese text [7]. Four finite automata perform the nominalization process: the synonymy automaton with 327 entries, the exception automaton with 4,223 exceptions, the adjective pattern automaton with 663 patterns, and the verb pattern automaton with 351 patterns. The rules for the derivation used by these automata were manually constructed following the descriptions found in the Aurélio Portuguese Dictionary [3].

Figure 2 shows the output for the sentence "Controle trabalhista exato adotado em equipe" ("Exact labor control adopted in team").

```
Controle controle 0 0 _SU
trabalhista trabalhista trabalho 0 _AJ
exato exato exatidao 0 _AJ
adotado adotar adocao adotante _AP
em em 0 0 _PR
equipe equipe 0 0 _SU
. . 0 0 _PN
```

Figure 2. FORMA and CHAMA output example for Brazilian Portuguese

The output per line in Figure 2 is organized as a set containing:

– the original word, (e.g., "adotado" ("adopted")),
– the lemma, (e.g., "adotar" ("to adopt")),

– the abstract noun (e.g., "trabalho" ("work"), "exatidão" ("accuracy") and "adoção" ("adoption")), when it exists, or zero, if there is no nominalization,
– the concrete noun (e.g., "adotante" ("who adopts")), when it exists, or zero, if there is no nominalization, and
– the part-of-speech tag (in Figure 2: _SU=noun, _AJ=adjective, _AP=participle, _PR=preposition, and _PN=pontuaction mark).

3.2 Binary Lexical Relations

BLRs [4] identify relationships between nominalized terms. These relationships capture phrasal cohesion mechanisms [8] like those that occur between subject and predicate, subject and object (direct or indirect), noun and adjective or verb and adverb. Such mechanisms reveal term dependences.

A BLR has the form *id*(*t*1,*t*2) where *id* is a relation identifier, and *t*1 and *t*2 are its arguments (nominalized terms).

There are three kinds of BLRs:

– **Classification**: where *id* is the equal sign, *t*1 is a subclass or an instance of *t*2, and *t*2 is a class. The classification BLR example =(dida,goalkeeper) may be extracted from the string the goalkeeper Dida.
– **Restriction**: where *id* is a preposition, *t*1 is a modifier and *t*2 is its head. The mapping of syntactic dependencies onto semantic relations [1], concerning the prepositions, is the purpose of the BLR restriction. The restriction BLR example of(quickness,team) may be extracted from the string the quick team.
– **Association**: where *id* is an event, *t*1 is a subject and *t*2 is a direct or indirect object. An association may be prepositioned or not. The prepositioned association BLR example travel.across(tourist,europe) may be extracted from the string the tourist traveled across Europe. The association BLR example training(coach,athlete) may be extracted from the string the coach trained the athlete.

We developed a tool named RELLEX that automatically extracts BLRs from a Brazilian Portuguese text. For more details about BLRs and BLR extraction rules see [4]. Those rules, and the nominalization process, are resources used to extract a unique BLR derived from different syntactic structures with the same semantics.

3.3 Evidence concept and descriptor weighting

Evidence is information that gives a strong reason for believing something or that proves something; evidences are signs, indications; something is evident if it is obvious [2, 3]. The evidence concept is crucial for TR+ Model that adopts descriptor weighting based on this concept, i.e., the weighting is not only based on the descriptor occurrence frequency. The descriptor representativeness depends, besides the frequency of occurrence, on the occurrence of phrasal cohesion mechanisms.

The evidence ($evd_{t,d}$) of a term *t* in a document *d* is:

$$evd_{t,d} = \frac{f_{t,d}}{2} + \sum_{r} f_{r,t,d} \tag{1}$$

where:
$f_{t,d}$ is the occurrence frequency of *t* in *d*, and
$f_{r,t,d}$ is the number of *BLR*s in *d* where *t* is an argument.

On the other hand, the evidence $evd_{r,d}$ of a BLR *r* in a document *d* is:

$$evd_{r,d} = f_{r,d}\,(evd_{t1,d} + evd_{t2,d}) \tag{2}$$

where:
$f_{r,d}$ is the occurrence frequency of *r* in *d*, and
$evd_{t1,d}$ and $evd_{t2,d}$ are the evidences of *t*1 and *t*2, respectively, and *t*1 and *t*2 are arguments of *r*.

We adopted the Okapi BM25 formula [10] without *IDF* factor (which did not contribute for improvements for TR+ Model). So, the weight $W_{i,d}$ based on the evidence concept for a descriptor *i* (a nominalized term or a BLR) in a document *d* is given by:

$$W_{i,d} = \frac{evd_{i,d}(k_1+1)}{k_1((1-b)+b\frac{DL_d}{AVDL})+evd_{i,d}} \quad (3)$$

where:
k_1 and b are parameters whose values are 1.2 and 0.75 respectively;
DL_d is the length of d and $AVDL$ is the average document length in the collection; and
$evd_{i,d}$ is the descriptor evidence ($evd_{t,d}$ for a term t or $evd_{r,d}$ for a BLR r).

In TR+ Model, query descriptors have their weight computed by the same formula used for documents. The relevance value $RV_{d,q}$ of a document d for a query q is given by:

$$RV_{d,q} = \sum_i \frac{W_{i,d}W_{i,q}}{s_i} \quad (4)$$

where:
i is a term or a BLR;
$W_{i,d}$ is the weight for descriptor i in d;
$W_{i,q}$ is the weight for descriptor i in q; and
$s_i = 2$, if i is a BLR with the same arguments but different relation identifiers in d and q, or
$s_i = 1$, if i is a term or a BLR with the same arguments and identifiers.

The document classification depends on the relevance values and the Boolean query formulation.

3.4 Boolean query and grouped classification of documents

A query q formulated by the user, in TR+ Model, is recognized as a text like a text document. A Boolean query qb, automatically derived from a query q, is formulated according to the following grammar (in EBNF formalism):

```
<qb> → [ <BLRDisj>  OR ] <TermConj>
<BLRDisj> → <r>  [ OR  <BLRDisj> ]
<TermConj> → (<TermDisj>  [ AND  <TermConj> ])
<TermDisj> → (η1(<w>) OR η2(<w>)) | (η1(<w>)) | (η2(<w>)) |
<r> → BLR
<w> → adjective | adverb | noun | verb
```

The elements OR and AND are respectively disjunction and conjunction Boolean operators. Let the string "restored painting" is a query q, then a corresponding Boolean query qb is:

"of(restoration, painting) " OR (("restoration" OR "restorer") AND ("painting")))

In the next step, the retrieved documents are classified in two groups:

– **Group I**: more relevant documents that fulfill the Boolean query conditions; and
– **Group II**: less relevant documents that do not completely fulfill the Boolean query conditions, but contain at least one query term.

In each of these groups, the documents are ranked in decreasing order of relevance value according to equation (4).

4 Collections and indexing files

We submitted only one run for the Portuguese Monolingual Ad Hoc Task at CLEF 2006: the trevd06. This run adopts the TR+ Model, i.e., uses terms and relationships in evidence.

Trevd06 uses manual query construction from topics, including title, description, and narrative. For each topic, a query in natural language was entered. Such text constituted the input of our system according to the strategy of TR+ Model (see Figure 1).

PT collections at CLEF 2006 were Publico95, Publico94, Folha95, and Folha94. Table 1 shows the amounts of descriptors (terms and BLRs) extracted from each collection.

Table 1. Terms and BLRs from each collection

collections	# of terms	# of classi-fications	# of restric-tions	# of asso-ciations	# of prep. associa-tions
Publico95	218763	987332	2604167	155246	174311
Publico94	211011	932750	2462727	144062	160737
Folha95	198343	665776	1798231	107573	103244
Folha94	191798	645769	1739184	103482	99802

Figure 3 shows the indexing file structure. The "term vocabulary" file is a string of terms t delimited by \0 and the "preposition vocabulary" file is a string of prepositions p delimited by \0. If a t is the T^{th} term in the vocabulary, then T is the term ID of t. The same strategy is applied to prepositions.

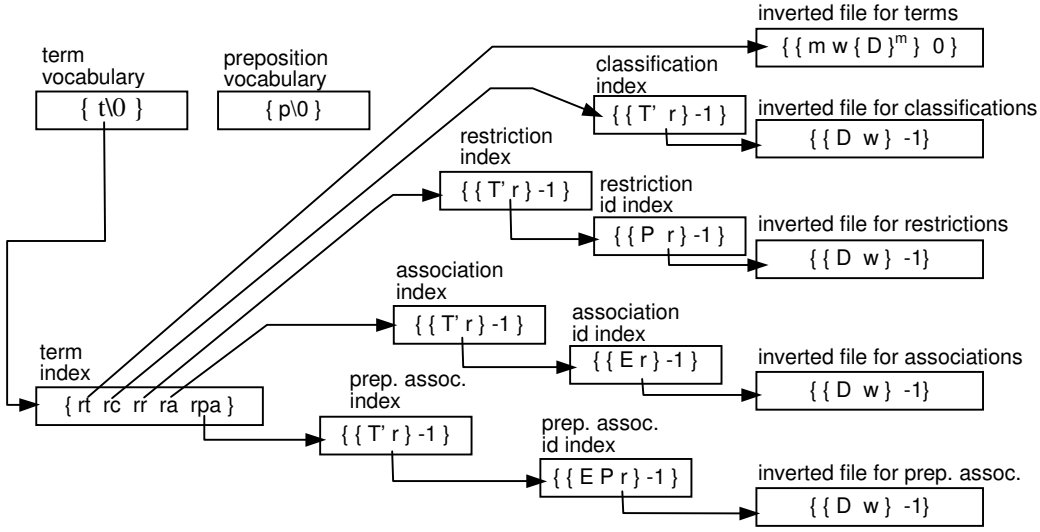

Figure 3. Indexing file structure

The T^{th} record in "term index" file has five offset positions for the term t: rt for the "inverted file for terms", rc for the "classification index" file, rr for the "restriction index" file, ra for the "association index" file, and rpa for "prepositioned association index" file. The default value is -1 if there is no corresponding offset position in those files.

The corresponding record for term t in "inverted file for terms" starts with m, indicating the size of the list of document IDs $\{ D \}^m$ where term t occurs with weight w. The inverted list for a term t consists of a sequence $m\ w\ \{ D \}^m$, which is repeated until a m value of 0 is found.

In "classification index" file, the corresponding record for term t has T' (a term ID of t') and the offset position r in the "inverted file for classifications" for the classification =(t,t'). The sequence T' r for term t (indicating classifications where t is the first argument) is repeated until a T' value of -1 is found. The same strategy is used for the "restriction index", "association index", and "prepositioned association index" files. In these cases, r is an offset position in "id index" files.

The corresponding record for the restriction p(t,t') in "restriction id index" file has P (a preposition ID of p) and the offset position r in the "inverted file for restrictions". The corresponding record for the association e(t,t') in "association id index" file has E (a term ID of e) and the offset position r in the "inverted file for associations". The corresponding record for the prepositioned association e.p(t,t') in "prepositioned association id index" file has E (a term ID of e), P (a preposition ID of p), and the offset position r in the "inverted file for prepositioned associations".

For the BLRs, the corresponding record in inverted files has D (a document ID of d), where the BLR occurs with weight w. The inverted list for a BLR consists of a sequence D w, which is repeated until a D value of -1 is found.

3.4 Difficulties

This is our first participation in CLEF. Our main goal was to obtain hands-on experience in the Ad Hoc Monolingual Track on text such as the PT collections. Our prior experience was indexing and searching smaller text collections using Brazilian Portuguese only.

The changes applied to a search engine for such task are not simple adjustments and the decision to participate was taken late. Some misplaces were verified during the indexation phase of trevd06. Our estimation is that at least 20% of the terms were not indexed due to programming mistakes.

The differences between Brazilian Portuguese and European one are another source of errors because our system was designed for the former but not for the European Portuguese. The following example explains this problem. While Figure 2 shows the output of our nominalization tool for a sentence in Brazilian Portuguese, Figure 4 shows the output for the same sentence ("Controlo laboral exacto adoptado em equipa") in European Portuguese.

```
Controlo controlar controle controlador _VB
laboral laboral laboralidade 0 _AJ
exacto exacto 0 0 _SU
adoptado adoptar adoptacao adoptador _AP
em em 0 0 _PR
equipa equipar equipamento equipador _VB
```

Figure 4. FORMA and CHAMA output example for European Portuguese

You should notice that the noun "Controlo" here is tagged as a verb (_VB), the adjective "exacto" as a noun (_SU), and the noun "equipa" as a verb. Nouns, like "laboralidade" and "adoptação", are generated erroneously due to lexical differences between the two languages. On the other hand, nouns, like "controle", "controlador", "equipamento", and the wrong noun "equipador", are generated due to incorrect tagging. These mistakes affected the indexing of terms and BLRs.

4 Results

Figure 5 presents interpolated recall vs average precision for the trevd06 run.

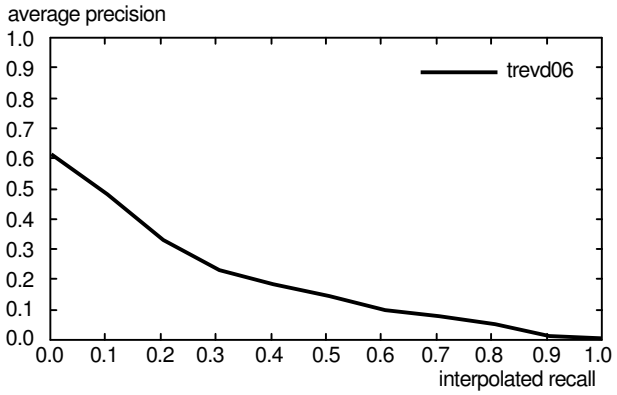

Figure 5. Interpolated recall vs average precision

There are 2566 relevant documents concerning the 50 topics of the Ad Hoc Monolingual Portuguese Track at CLEF2006. Our run retrieved 1298 relevant documents. Some of the results are shown here:

- Average Precision: 18,04%
- R-Precision: 22,85%
- Precision at 5 docs: 40,0%
- Precision at 10 docs: 35,0%

5 Conclusion

Indeed, under the conditions of participation in Ad Hoc Monolingual Portuguese Track at CLEF2006, we could consider that the results were reasonable and the experience with European Portuguese and largest colletions was valid.

There are two immediate works concerning this experience:

– the correction of the indexing errors and the analysis of the impact on retrieving results, and
– the adaptation of our tagging and nominalization tools concerning European Portuguese.

We must decide on two work directions: (i) to use specialized tools for each language type or (ii) to create a generic tool for text pre-processing. An alternative that will be considered is to transform variations of words (like "exato" and "exacto" ("exact")) into a common form before the indexing phase.

References

[1] Gamallo, P.; Gonzalez, M.; Agustini, A.; Lopes, G; Lima, V. L. S. de. Mapping Syntactic Dependencies onto Semantic Relations. ECAI'02, Workshop on Natural Language Processing and Machine Learning for Ontology Engineering, Lyon, France, 2002. p.15-22.

[2] Crowther, J. (ed.). Oxford Advanced Learner's Dictionary of Current English. New York: Oxford University Press, 1995. 1,430 p.

[3] Ferreira, A. B. H. Dicionário Aurélio Eletrônico – Século XXI. Nova Fronteira S.A., Rio de Janeiro, 1999.

[4] Gonzalez, M.; Lima, V. L. S. de; Lima, J. V. de. Binary Lexical Relations for Text Representation in Information Retrieval. *10th Int. Conf on Applications of NL to Inf. Systems*, NLDB, 2005. Springer-Verlag, LNCS 3513, 2005. p.21-31.

[5] Gonzalez, M. Termos e Relacionamentos em Evidência na Recuperação de Informação. PhD thesis, Instituto de Informática, UFRGS, 2005.

[6] Gonzalez, M.; Lima, V. L. S. de; Lima, J. V. de. *7th Comp. Ling. and Intel. Text Processing* - CICling, 2006. Springer-Verlag, LNCS 3878, 2006. p.394-405.

[7] Gonzalez, M.; Lima, V. L. S. de; Lima, J. V. de. *7º Encontro para Proc. Comp. da Língua Portuguesa Escrita e Falada* - PROPOR, 2006. Springer-Verlag, LNCS 3960, 2006. p.100-109.

[8] Mira Mateus, M. H.; Brito, A. M.; Duarte, I.; Faria, I. H. Gramática da Língua Portuguesa. Lisboa: Ed. Caminho, 2003.

[9] Perini, Mário A. *Para uma Nova Gramática do Português*. São Paulo: Ed. Ática, 2000.

[10] Robertson, S. E.; Walker, S. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. *17th Annual International ACM SIGIR conference on research and development in IR*, 1994. Proceedings, p.232-241.

[11] Ziviani, N. Text Operations. In: Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*. New York : ACM Press, 1999.