

Categorization-Driven Cross-Language Retrieval of Medical Information

Hermes R. Freitas-Junior¹ Berthier Ribeiro-Neto^{1,2}
Rodrigo F. Vale¹ Alberto H. F. Laender² Luciano R. S. Lima^{2,3}

¹ Akwan Information Technologies
Av. Antônio Abraão Caram, 430 - Zip 31275-000
Belo Horizonte, MG, Brazil
{hermes,rodrigov,berthier}@akwan.com.br

² Department of Computer Science
Federal University of Minas Gerais
Av. Antônio Carlos, 6627 - Zip 31270-901
Belo Horizonte, MG, Brazil
{berthier,laender,luciano}@dcc.ufmg.br

³ The Sarah Network of Hospitals for Rehabilitation
Av. Amazonas, 5953 - Zip 30510-000
Belo Horizonte, MG, Brazil
lucianors@sarah.br

March 11, 2005

Abstract

The Web has become a large repository of documents (or pages) written in many different languages. In this context, traditional information retrieval (IR) techniques cannot be used whenever the user query and the documents being retrieved are in different languages. To address this problem, new cross-language information retrieval (CLIR) techniques have been proposed. In this work, we describe a method for cross-language retrieval of medical information. This method combines query terms and related medical concepts obtained automatically through a categorization procedure. The medical concepts are used to create a linguistic abstraction that allows retrieving information in a language-independent way, minimizing linguistic problems such as polysemy. To evaluate our method, we carried out experiments using the OHSUMED test collection, whose documents are written in English, with queries expressed in Portuguese, Spanish, and French. The results indicate that our cross-language retrieval method is as effective as a standard vector space model algorithm operating on queries and documents in the same language. Further, our results improve previous results in the literature.

1 Introduction

With the rapid expansion of the Web, large amounts of information, written in several languages, have become available. While the English language is predominant, it is not absolute. Languages with smaller representation, such as Portuguese and Spanish, are in expansion. In this context, the information made available in one language is frequently accessible to native speakers of a second language. However, while people are at ease with the reading of a document written in a foreign language, they often find it hard to speak or write in that language. A popular proposal to circumvent this problem is to enrich information retrieval (IR) systems with cross-language retrieval capabilities. In this case, queries written in a first language are used to retrieve information made available in a second language; a new reality that presents its own challenges.

Cross-language information retrieval (CLIR) techniques are useful because they considerably simplify the query formulation process, especially in those situations in which the user is not fully fluent in the language used to write the documents of the collection. Additionally, CLIR techniques allow the use of a single query to retrieve documents written in distinct languages, avoiding the need to formulate the query in various languages.

Current IR systems do not provide support for CLIR, and are, usually, unable to properly handle the problems that arise when queries and documents are in distinct languages. In traditional IR models, the index term is the basic information unit. Thus, queries and documents are expressed as a set of terms [Baeza-Yates & Ribeiro-Neto, 1999]. This term-based paradigm is quite convenient because it is simple and easy to implement and use. However, term-based retrieval makes it difficult to represent the information semantics in many situations. For instance, in CLIR environments, the set of terms that represents specific information in one language might not have the same meaning in a second language. If a more representative information unit (other than terms) is used, the semantic representation of documents and queries can be considerably improved. This information unit, generally called concept, can be obtained from taxonomies or thesauri that aggregate human knowledge.

In the medical domain, such taxonomies have been widely used and are obtained from controlled vocabularies [Cimino, 1995]. They are used to categorize information according to medical concepts defined in the controlled vocabularies. This categorization process is very important because it helps institutions to organize their medical records, which are produced in abundance. These records are useful for gathering statistics on health care services and to support billing.

In this paper, we describe a CLIR method applied to the medical domain. This method uses a categorization algorithm to map free-text documents and queries to medical concepts in fully automatic fashion, therefore providing a means to represent and recognize such concepts in different languages. To automatically categorize medical information, we adopt a categorizer based on the *HiMeD* model [Lima et al., 1998, Ribeiro-Neto et al., 2001]. This automatic categorizer is responsible for identifying, extracting, and encoding medical information contained in the user query and in the documents of the collection, according to the concepts of the Medical Subject Headings (MeSH) [mes, 2000]. This process plays a central role in our method because it makes it possible to construct semantic units that minimize linguistic problems (e.g., polysemy) generated by the use of textual queries.

The medical concepts acquired, which represent a new source of evidential knowledge, and the original query terms need to be properly combined so that the retrieval process can return relevant information. In this context, traditional ranking algorithms derived from the vector model [Baeza-Yates & Ribeiro-Neto, 1999, Salton, 1968] are difficult to use, because they were not designed to work with new sources of evidence. To address this issue, we adopt the formal framework of belief networks [Baeza-Yates & Ribeiro-Neto, 1999, Ribeiro-Neto & Muntz, 1996]. In our method, this framework provides support to a ranking fusion model

we introduce, which is designed to combine the rankings generated by both sources of evidence, the medical concepts and the free-text terms. We also use a generic dictionary to translate common query terms. This dictionary, used as a complementary component, makes possible the translation of terms not present in the MeSH vocabulary.

To demonstrate the effectiveness of our method, we carry out a set of experiments using OHSUMED [Hersh et al., 1994], a test collection based on a subset of MEDLINE¹. Our experiments show that our CLIR method is equally effective for textual queries in Spanish, Portuguese, and French, when compared with the results generated by a standard IR method that takes queries and documents in English. Furthermore, our results are superior to previous results in the literature, obtained using the same experimental environment we adopt. The main ideas discussed in this paper have been implemented in a experimental medical search engine we developed.

The remainder of the paper is organized as follows. Section 2 discusses related work. Our CLIR method is described in Section 3. Section 4 discusses the fusion ranking model. Experimental results are presented in Section 5. Finally, Section 6 presents our conclusions.

2 Related Work

Salton [Salton, 1970] was the first to demonstrate that IR techniques could be used in a multilingual environment. His system was based on the vector model and used a manually created thesaurus. This thesaurus basically worked as a dictionary that was used to translate the query terms. This pioneer work showed how traditional IR techniques could be extended to be applied in a multilingual context.

Currently, the major approaches for CLIR use distinct techniques. The most simple ones are based on multilingual dictionaries [Ballesteros & Croft, 1996] that translate terms from one language to another. The most complex approaches use corpus-based techniques supported by parallel or comparable collections. Corpus-based techniques use empirical associations extracted from bilingually aligned documents. Examples of these techniques include Latent Semantic Indexing [Dumais et al., 1996], the Generalized Vector Space Model [Yang et al., 1998], and Statistical Machine Translation [Aljlayl & Frieder, 2001, Federico & Bertoldi, 2002, Gao et al., 2001, Nie et al., 1999]. Some works attempt to combine features of various approaches. For instance, Resnik et al. [Resnik et al., 2001] use evidence from dictionary-based and the corpus-based approaches in their work. Other works in-

¹<http://www.ncbi.nlm.nih.gov/pubmed>

investigate how human interaction can improve the cross-language retrieval [Oard, 2000, Ogden & Davis, 2000].

New and interesting techniques also have been proposed. The most promising [Gollins & Sanderson, 2001] try to minimize problems of the simple word-by-word translations creating intermediate (or pivot) languages representation, an idea explored by our work. The work in [Lavrenko et al., 2002] proposes a formal model that, instead of relying on either query or document translation, uses techniques of disambiguation and query expansion to rank documents with regard to a query in another language.

In the medical domain context, a major reference is the work by Eichmann et al. [Eichmann et al., 1998]. This work is an example of a thesaurus-based approach whose main idea consists in translating the user's queries through the UMLS Metathesaurus [uml, 2001]. The UMLS Metathesaurus was created by the National Library of Medicine and includes many controlled vocabularies of which the most important one is the Medical Subject Headings (MeSH). In their work, Eichmann and his colleagues used the OHSUMED test collection [Hersh et al., 1994], which includes a set of 106 test queries formulated in English. For evaluating their CLIR method, these original queries were first translated into Spanish and French, and then translated back to English by their automatic method. The original English queries were used as the baseline and considered as the upper bound performance for the experiment. The performance achieved for the Spanish and French queries was about 60-70% of the corresponding baseline. This result was quite interesting because it shows that the performance achieved by using a thesaurus as a specialized dictionary is not much better when compared with the use of generic dictionaries.

Another interesting CLIR approach applied to the medical domain, proposed by Rassinoux et al. [Rassinoux et al., 1998], use ontologies to store the information in an intermediary format that makes it possible to retrieve documents in a language-independent way. To map free-texts into ontology concepts it uses NLP techniques. Similarly, Zweigengaum et al. [Zweigenbaum et al., 1995] propose an architecture to manage medical information contained in Patient Discharge Summaries (PDS). The architecture is based on the MENELAS system [Whittemore, 1994] that stores additional PDS related data. This additional data, based on ICD codes [cid, 1980], makes it possible to retrieve information in a language-independent way. The *Analysis System* component analyzes the PDS and stores them as a set of conceptual language-independent representations. Our work is similar to these because we also use an intermediary and language-independent representation of the document contents. However, we use the framework based of belief networks to allow combining multiple sources of evidence into a single ranking, a proce-

ture that improves the quality of the retrieval.

Finally, some initiatives on the Web, such as *CliniWeb* [Hersh et al., 1996], which is supported by the *SAPHIRE* [Hersh & Donohoe, 1998] system, and the *much.more* project² should be mentioned. *CliniWeb* catalogues about 10,000 Web pages from medical sites. The *SAPHIRE* system is used to associate the query terms with UMLS Metathesaurus concepts. Thus, once each page stored in the *CliniWeb* database has MeSH terms manually assigned to it, it is possible to retrieve information in a language-independent way. The *much.more* project is more recent and aims to develop technologies that will result in a prototype system for cross-language information organization and retrieval for the medical domain. The work is focussed on techniques such as query disambiguation [Steffen et al., 2003], semantic annotation [Sacaleanu et al., 2003] and automatic thesaurus construction [Widdows, 2003].

We can contextualize our CLIR method based on the framework proposed by Oard in [Oard, 1997]. Oard presents two broad categories to classify the main CLIR approaches: those based on *controlled vocabularies* and those based on *free-text* methods. The later can be split in two other categories based on parallel or comparable document collections (*corpus-based*) and based on dictionaries or NLP tools (*knowledge-based*). Our method can be classified as a *controlled vocabulary* technique with some features from *knowledge-based* techniques, because we also use a generic dictionary. The *controlled vocabulary* technique presents some challenges that tend to reduce its effectiveness [Oard, 1997]. The first problem is that the vocabulary terms must be assigned to each document in the collection. The second is the difficulty to train users to use the correct terms of the vocabulary and to exploit its relationships.

Our method attempts to mitigate these problems by providing a mechanism for the correct acquisition and representation of the vocabulary terms. We do not work with terms but with medical concepts (derived from the MeSH vocabulary), which aggregate more semantic information than terms. The concepts are automatically assigned to each document in the collection and to the user's query. This assignment is performed by an automatic categorizer that is responsible to acquire the medical concepts. In addition, we introduce the fusion ranking model that allows combining evidence derived from index terms with information derived from medical concepts to yield improved retrieval performance.

²<http://muchmore.dfki.de>

3 The Method

We propose a new CLIR approach for medical collections that effectively provides retrieval performance as good as that obtained with standard (one language) IR systems. Our idea is to use hierarchical taxonomies of concepts, specified in various languages, to assist with the cross-language retrieval process. We focus on the medical domain, where such taxonomies are well established. Our approach is distinct because we consider that the medical documents have been previously categorized in the taxonomy considered. This allows using information on the medical concepts associated with each document to considerably improve the quality of our results. Most important, the categorization procedure is done in fully automatic fashion using algorithms described in [Lima et al., 1998, Ribeiro-Neto et al., 2001].

3.1 Overview

Acquisition and representation of medical knowledge play a central role in our method. The acquisition process is performed by an automatic categorizer developed according to [Lima et al., 1998, Ribeiro-Neto et al., 2001], while the representation process is performed by a ranking fusion model we discuss here. The categorizer is used to encode the documents of the collection using MeSH concepts [mes, 2000]. The MeSH medical concepts are quite important in our approach because they allow retrieval of information while minimizing linguistic problems. A generic dictionary is also used as a complementary technique. The dictionary is used to expand the user query with translation of common terms not captured by the MeSH vocabulary.

One important feature of our method is the flexibility it provides in the query formulation process. In fact, the user can express his information need as a set of terms, as medical concepts or as a union of both, as done in Web directories. If the user expresses his query using terms, the automatic categorizer finds the related medical concepts and the dictionary translates the common terms. The final query is processed as a union of original terms, of medical concepts and of translated terms. Figure 1 illustrates an example in which q_o , q_c and q_t are the original query, the related concepts and the translated terms, respectively. The original query q_o is stated in Portuguese. The categorizer associates the medical concepts labeled D004358 and D001943 to the query. These two concepts form the set q_c . The terms "chemotherapy", "advanced", and "suck" form the set q_t of translated terms. If the user query contains only references to medical concepts, then the final query contains only the query component q_c .

⇒ **Insert Figure 1 here.**

3.2 The MeSH Vocabulary

The Medical Subject Headings [mes, 2000] was created by the National Library of Medicine to be the controlled vocabulary used for indexing articles, for cataloging books and other holdings, and for searching MeSH-indexed databases, including MEDLINE. MeSH terminology provides a consistent way to retrieve information that may use different terminology for the same concepts. MeSH organizes its descriptors in a hierarchical structure so that broad searches will find articles indexed more narrowly.

The MeSH vocabulary is continually updated by subject specialists in various areas. Each year hundreds of new concepts are added and thousands of modifications are made. The 2001 MeSH version, used in our work, includes more than 19,000 main headings, 115,000 Supplementary Concept Records (formerly Supplementary Chemical Records), and an entry vocabulary of over 300,000 terms.

Basically, the MeSH vocabulary consists of a set of terms or subject headings that are organized in an hierarchical structure. At the most general level of the hierarchical structure, one finds very broad headings such as *Anatomy* or *Mental Disorders*. At more narrow levels, one finds more specific headings such as *Ankle*, *Conduct Disorder* and *Calcineurin*. There are more than 19,000 main headings in MeSH. In addition to these headings, there are 103,500 headings called Supplementary Chemical Records within a separate chemical thesaurus.

3.3 The Automatic Categorizer

The automatic categorizer is used to categorize medical information. It is the component of our method responsible for the acquisition of medical knowledge. The extraction of concepts provided by the categorizer makes possible the construction of an index that combines terms and medical concepts. This mixed index provides support to the ranking fusion model we later discuss.

Our automatic categorizer is based on the *HiMeD* model [Lima et al., 1998, Ribeiro-Neto et al., 2001], which defines components and rules that deal with synonyms, acronyms and morphologic variations. Figure 2 exemplify the document classifying process. The algorithm operates in a fully automatic mode and requires no supervision or training data. The *HiMeD* model was created to assign ICD codes to medical documents. In here, it was adapted to support MeSH codes. When compared to similar models, the *HiMeD* model presents superior performance [Lima et al., 1998], because it is able to take advantage of the hierarchical structure of the thesauri.

⇒ **Insert Figure 2 here.**

Given a set of terms, the categorizer returns a set of related medical concepts. The categorizer can be defined as a function $categ(T)$ that, given an input text T , returns a set of concepts c_i . The text T corresponds to a query q whenever the categorizer is applied to the user's query, and to a document d_j whenever the categorizer is applied to the documents of the collection. Thus, given a set C of medical concepts extracted from the user's query, we immediately are able to identify subsets of related documents. These documents are those whose automatic categorization led to an association with the concepts in C .

3.4 The Generic Dictionary

The generic dictionary is used to translate common query terms, prior to retrieval of documents. The translated terms are used to expand the original query. We can formalize the use of this generic dictionary as follows. Let the user's query q be defined by a set of terms $q = \{k_1, k_2, \dots\}$. A function $dic(q, I_L)$, for a language I_L , is defined. This function returns $K_t = \{k'_1, k'_2, \dots, k'_i, \dots\}$, where each k'_i is a translation of a corresponding term k_i , in the language I_L . We can define a translated query $q_t = dic(q, I_1) \cup dic(q, I_2) \cup \dots \cup dic(q, I_n)$ as the union of all possible translations of the query q in various languages. Thus, the expanded query q' is defined as $q' = q \cup q_t$.

The dictionary technique is frequently used in CLIR approaches. However, the use of a dictionary presents problems that decrease the retrieval performance of the system when the technique is used in isolation. Basically, these problems are related to the ambiguous nature of the terms, which can have distinct meanings depending on the context used (technically, this is referred to as polysemy). We do not concern ourselves with these problems because the main terms that describe medical topics such as diseases, anatomy or drugs are covered by the categorizer.

4 The Ranking Fusion Model

The ranking fusion model is used in our method to combine evidence derived from medical concepts with evidence provided by query terms. It is based on the belief network model, which we now discuss.

4.1 Belief Network Model

The belief network model, introduced in [Ribeiro-Neto & Muntz, 1996], is based on Bayesian networks. Bayesian networks are a graphical formalism for

explicitly representing independencies among the variables of a joint probability distribution. The probability distribution is represented through a directed acyclic graph whose nodes represent the random variables of the distribution. The relationships among these variables are modeled as directed edges (in the graph) which represent causal dependencies among the linked variables (or nodes)³. The strengths of these dependencies are expressed by conditional probabilities.

In a traditional information retrieval system, documents are indexed by keywords (or terms) to facilitate the automatic retrieval of information. We interpret the set of all keywords as the universe of discourse \mathcal{U} , which we take as our sample space (as in [Wong & Yao, 1995]). Let t be the total number of keywords in a collection. Then we can define,

t : the total number of keywords

k_i : a keyword.

$\mathcal{U} = \{k_1, k_2, \dots, k_t\}$: the whole set of keywords. \mathcal{U} is interpreted as a concept space (our sample space). Each keyword k_i is interpreted as an elementary concept in the space \mathcal{U} .

$u \subseteq \mathcal{U}$: the concept u is a subset of \mathcal{U} composed of elementary concepts⁴.

Associated with each keyword k_i , we define a binary random variable which is also denoted by k_i . This variable is 1 to indicate that the keyword belongs to a given concept. Consider, for instance, a collection with t keywords in which a document d is represented as a concept $d = \{k_1, k_2, \dots, k_t\}$, $d \subseteq \mathcal{U}$, where each k_i is 1 if the keyword is in the document d , and 0 otherwise. Analogously, a query q can be represented as a concept $q = \{k'_1, k'_2, \dots, k'_t\}$, $q \subseteq \mathcal{U}$, where $k'_i \in \{0, 1\}$. To allow referring to the state of each variable k_i , we define a function $g_i(u)$ that returns the value of the variable k_i according to the concept u . The function $g_i(u)$ is 1 if $k_i \in u$, and 0 if $k_i \notin u$.

Given the concept space \mathcal{U} , it is natural to model queries and documents as concepts in \mathcal{U} . As a result, queries and documents are treated analogously. This symmetry induces the belief network of Figure 3.

\Rightarrow **Insert Figure 3 here.**

In this network, each node d_j models a document in the collection and the node q models the user query. With the node q is associated a binary random

³Although an edge linking a node Y to a node X frequently is used to express that Y causes X , this interpretation of edges in Bayesian networks is not the only one possible.

⁴This set-theoretic vision of concepts permits to reason with the logical notions of conjunction, disjunction, negation, and implication as operations of intersection, union, complementation, and inclusion [Wong & Yao, 1991].

variable which is also denoted by q . This variable is 1 (one) to indicate that q completely covers the sample space \mathcal{U} . The ranking computation is based on interpreting the similarity between a document d_j and the query q as a coverage relationship between the concepts d_j and q . To quantify the degree of coverage of the concept d_j , given the concept q , we use the probability $P(d_j|q)$. Applying Bayes' law, we can write $P(d_j|q) = P(d_j \wedge q)/P(q)$. Since $P(q)$ is constant for all documents, it is sufficient to compute $P(d_j \wedge q)$, which can be done through the expression $P(d_j \wedge q) = \sum_u P(d_j, q|u)P(u)$. Let η be a normalizing constant (as used in [Pearl, 1988]). Then,

$$P(d_j|q) = \eta \sum_u P(d_j|u) P(q|u) P(u) \quad (1)$$

which is the generic expression for the rank of a document d_j with regard to the query q .

Through proper specification of the probabilities $P(d_j|u)$ and $P(q|u)$, we can make the belief network model rank documents in the same ordering as the vector space model [Ribeiro-Neto & Muntz, 1996].

4.2 Ranking Fusion

The ranking fusion model is based on the belief network model and is used to formally represent the user query and pieces of evidence associated with it (generated by the translation process). These pieces of evidence are characterized by terms and MeSH medical concepts, as we now discuss.

Figure 4 illustrates the network for our ranking fusion model. The left hand side of the network is the network illustrated in Figure 4 (that considers only index terms, also called keywords) and the right hand side represents evidence derived from medical concepts c_j . In the network on left side, the user's query and the documents are modeled by the nodes q_k and d_{k_j} , respectively, where the k stands for keywords. In the right hand side of the network, we represent the user's query and the documents by the nodes q_c e d_{c_j} , respectively. With each medical concept c_j , we associate a set of documents represented as d_{c_j} nodes. These documents are the ones that are associated with the concept c_j by the automatic categorizer. Also, with the user's query we associate a set of concepts c_j . This set of concepts leads to a representation of the user query that we model as q_c . An extra node q is added at the top to combine all sources of query related evidences. Extra nodes d_j are also added at the bottom side to represent the combination of document related evidences. To combine the sources of evidences, we use a disjunctive operator.

⇒ **Insert Figure 4 here.**

The ranking generated by our network can be calculated through the probability $P(d_j|q)$. By the Bayes rules, we calculate this ranking as follows

$$\begin{aligned}
P(d_j|q) &= P(d_j \wedge q)/P(q) \\
&= \eta \sum_{u,v} P(d_j | u, v) P(q | u, v) P(u) P(v) \\
&= \eta \sum_{u,v} [1 - (1 - P(d_{k_j} | u))(1 - P(d_{c_j} | v))] \times \\
&\quad [1 - (1 - P(q_k | u))(1 - P(q_c | v))] \times P(u) P(v)
\end{aligned} \tag{2}$$

where u represents a state of the set of k_i nodes, v represents a state of the set of c_i nodes, $P(d_{k_j} | u)$ and $P(d_{c_j} | v)$ define the probabilities that the document d_j is related to a set of terms and to a set of medical concepts, respectively. In a similar way, $P(q_k | u)$ and $P(q_c | v)$ define the probabilities that the query q is related to a set of terms and to a set of medical concepts, respectively. The normalization parameter η and the prior probabilities $P(u)$ and $P(v)$ work as constants in our model and do not affect the final ranking.

To simplify the computation, we consider the influence of all terms that compose q_k together. This is accomplished by defining

$$P(q_k | u) = \begin{cases} 1 & \text{if } \forall_i, g_i(q_k) = g_i(u) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $g_i(u)$ is a function that returns the state of the variable k_i according to u . That is, the node q_k is turned on for the single state u in which the variables k_i that are on are exactly those that compose q_k . To express this, we define

$$u = u_q \quad \text{if } \forall_i, g_i(u) = g_i(q_k) \tag{4}$$

For q_c , we can define analogously,

$$P(q_c | v) = \begin{cases} 1 & \text{if } \forall_i, g_i(q_c) = g_i(v) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

and

$$v = v_q \quad \text{if } \forall_i, g_i(v) = g_i(q_c) \tag{6}$$

where q_c is composed of the concepts associated (by our categorizer) with the user query. By substituting Equations 3, 4, 5, 6 into Equation 2, we obtain

$$P(d_j|q) = \eta [1 - (1 - P(d_{k_j} | u_q))(1 - P(d_{c_j} | v_q))] \tag{7}$$

because $P(u_q)$ e $P(v_q)$ are constants.

Equation 7 provides a rather simple and fast to compute equation. Yet, one that is effective as we later discuss in our empirical results.

Let N be the total number of documents in the collection; n_i and n_l be the number of documents that make reference to the term k_i and to the concept c_l , respectively; and $freq_{i,j}$ and $freq_{l,j}$ be the frequencies of occurrence of the term k_i and of the concept c_l in the document d_j , respectively. The factor $\max freq_{i,j}$ ($\max freq_{l,j}$) is the maximum frequency of occurrence among all terms (concepts) within the document d_j . The factor $\max \log N/n_i$ ($\max \log N/n_l$) is the maximum *idf*, inverse document frequency, among all terms (concepts) in the collection. These are used as normalization factors to fit our probabilistic framework. Our setup provides a variation of the *tf-idf* weighting schemes used in the vector space model, as follows.

We define

$$\begin{aligned}
 P(d_{k_j} | u_q) &= \sum_{k_i | g_i(u_q)=1} \left(\frac{freq_{i,j}}{\max freq_{i,j}} \right) \left(\frac{(\log N/n_i)^2}{\max \log N/n_i} \right) \\
 P(d_{c_j} | v_q) &= \sum_{c_l | g_l(v_q)=1} \left(\frac{freq_{l,j}}{\max freq_{l,j}} \right) \left(\frac{(\log N/n_l)^2}{\max \log N/n_l} \right)
 \end{aligned}
 \tag{8}$$

Notice that our *tf-idf* scheme does not normalize the ranking with the norm of the document. This weighting schema, named *nfx.bfx* by Salton and Buckley [Salton & Buckley, 1988], was used because it is appropriate for homogeneous collections composed of short technical documents, like our test collection. Further, it allows a smoother combination of evidences (q_k and q_c) in the formal ranking formula. To implement the cosine formula of the vector model [Salton, 1968], all we have to do is normalize Equation 8 using the norm of the document.

After definition of the various conditional probabilities, we apply Equation 7 to generate a ranking of documents with regard to a query q . This ranking combines evidence from index terms with evidence from medical concepts. In our cross-language retrieval approach, we take a query in a language I_L and translate it into English. This translation process leads to a set of index terms in English and to a set of related medical concepts.

5 Experiments

The experiments carried out to evaluate our CLIR method are very similar to the ones presented by Eichmann et. al in [Eichmann et al., 1998]. We used the same OHSUMED test collection [Hersh et al., 1994] and the same set

of English queries. These queries were translated into Spanish, Portuguese, and French by medical professionals fluent in these languages. To establish a baseline for reference, we processed the original English queries by applying the standard vector space model.

In the following subsections, we describe the OHSUMED test collection, the performance measures used, how the experiments were carried out, and the results achieved.

5.1 The OHSUMED Test Collection

The OHSUMED test collection [Hersh et al., 1994] is composed of a set of 348,566 medical documents extracted from MEDLINE⁵. The collection includes English language documents which are structured according to the following seven fields: id, title, source, authors, MeSH terms, publication type, and abstract. As in [Eichmann et al., 1998], we use only the subset of 233,445 (67%) documents that contain the abstract field. Figure 5 illustrates an example of an OHSUMED document.

⇒ **Insert Figure 5 here.**

The OHSUMED collection includes a set of 106 text queries written in English. In our experimentation, we use the 101 queries for which there are relevant documents assigned. Figure 6 shows examples of two OHSUMED queries.

⇒ **Insert Figure 6 here.**

5.2 Retrieval Performance Measures

We used the same retrieval performance measures adopted by Eichmann et al. in [Eichmann et al., 1998]. The first measure is the classic $AvgP_{11}$ or the average precision at 11 recall levels. The second measure computes exact precision scores for the top n documents (called Top_nP) retrieved. Similar to [Eichmann et al., 1998], we are interested in the top 10 ranked documents.

5.3 Experimental Setup

The first step in our experiments was to determine a baseline to be used as reference. This was done by running the original English queries with the standard vector space model (cosine formula with $tf - idf$ weights). The baseline simulates the usual retrieval environment in which queries and documents are in the same language. Our purpose was to evaluate how

⁵<http://www.ncbi.nlm.nih.gov/PubMed/>

close to this baseline our cross-language retrieval approach can reach. Thus, we compared the results of our method for the three languages considered, Portuguese, Spanish, and French, with the baseline.

To do this comparison, we used three basic query formulations: q_o , which represents a query in one of the three testing languages, q_c , which represents a query formed only by medical concepts automatically extracted from query q_o , and q_t which represents a query formed by English keywords translated from q_o with the help of the generic dictionary. These three basic queries were combined into four other queries: $q_{oc} = q_o \cup q_c$; $q_{ot} = q_o \cup q_t$; $q_{ct} = q_c \cup q_t$ and $q_{oct} = q_o \cup q_c \cup q_t$.

Figure 7 illustrates the basic queries generated by a translation to Portuguese of the English query "chemotherapy advanced for advanced metastatic breast cancer". We observe that the English word *breast* was initially translated to Portuguese as *mama*, which is correct. However, when we used the generic vocabulary to translate back the work *mama*, we obtained *suck* as the translation, which is incorrect. The mistake is difficult to avoid. In our method, the effects of such mistakes on the results are diminished due to the use of related medical concepts (in Figure 7, the MeSH concept D001943 retains the semantics of *breast cancer*).

⇒ **Insert Figure 7 here.**

We ran our CLIR method for each one of seven possible query formulations (i.e., q_o , q_c , q_t , q_{oc} , q_{ot} , q_{ct} , q_{oct}) and for the three testing languages we have considered.

5.4 Results and Analysis

Table 1 presents the $AvgP_{11}$ figures obtained by our baseline. In this table, q_o represents the original English query in English (composed of index terms), q_c represents the medical concepts extracted from query q_o , and q_{oc} represents the combination of both in our fusion network model. We compare results for the cosine (*cos*) formula of the standard vector model [Salton, 1968] and for our the *nfx.bfx* ranking formula of Equation 8.

⇒ **Insert Table 1 here.**

For the original query formulation q_o , we notice that the cosine formula is slightly superior. Thus, it is used as the baseline in the remaining of our experiments. We also show the results obtained when we do the retrieval using medical concepts (i.e., the query formulation q_c) obtained directly (and automatically) from the query q_o . We observe that the results are 5-10% inferior to our baseline (i.e., q_o with *cos*) and that the *nfx.bfx* schema yields better results. Finally, the row q_{oc} displays results generated by the combination of index terms and medical concepts in our fusion network model. In this

case, the schema *nfx.bfx* leads to results that are roughly 8% better than our baseline. This is an interesting result because we are not doing cross-language retrieval yet. That is, we are processing queries and documents in English but are enriching the original query formulation with medical concepts generated automatically. The inclusion of this new type of evidence in our network model yields improved retrieval performance.

In the second phase of our experiments, we calculated the $AvgP_{11}$ of our CLIR method considering the queries formulated in Portuguese, Spanish, and French. That is, we consider that the original q_o has been translated to Portuguese, Spanish, and French (and is still referred to as q_o). The documents retrieved are in English. Table 2 presents our results for seven possible query formulations. The values in parenthesis correspond to the percentage of the baseline, in terms of average precision, that was achieved. When only terms are used (query formulations q_o , q_t and q_{ot}), we observe that the best retrieval performance achieved is roughly 50% of the baseline in all languages tested. When medical concepts are used in isolation (q_c query formulation), the results are at least 83% of the baseline (for French) and can be as high as 98% (for Spanish). When we use combined query formulations, the results keep improving until we reach the q_{oct} combination that yields the highest results. For Portuguese and Spanish, the q_{oct} combination (done in our ranking fusion model) yields results as good as the baseline. For French, the results are only slightly inferior.

⇒ **Insert Table 2 here.**

Table 3 details recall and precision figures for the query combination q_{oct} . We notice that at 10% recall, precision figures are very close or superior to the baseline. In Table 4, we show exact precision scores (Top_nP) for the top 5, 10, 15, 20, 30 and 50 documents. For the top 10 documents, our method achieves its best results, overcoming the baseline in all languages tested.

⇒ **Insert Table 3 here.**

⇒ **Insert Table 4 here.**

Finally, we point out that our results improve considerably previous results in the literature. For instance, Eichmann et al. present in [Eichmann et al., 1998] results based on an identical experimentation environment. For queries in Spanish and French, they achieve $AvgP_{11}$ scores of 71% and 61%, respectively, when compared to the baseline. Our results are quite superior and achieve $AvgP_{11}$ scores of 101% and 94% for Spanish and French, respectively.

6 Conclusions

In this work, we presented a CLIR method specialized in the medical domain. The method is supported by the MeSH controlled vocabulary, which is used to identify medical concepts of relevance to the cross-language retrieval process. These medical concepts are used as semantic units, free of linguistic barriers.

In our method, medical concepts are associated with queries and documents in fully automatic fashion. This is accomplished with the use of an efficient categorizer developed according to [Lima et al., 1998, Ribeiro-Neto et al., 2001]. To combine evidence from index terms and medical concepts, we defined a fusion ranking model based on belief networks.

We evaluated our method using the OHSUMED test collection, a subset of MEDLINE. The original English queries were translated to Spanish, Portuguese and French by medical professionals doctors fluent in these languages. We then ran our CLIR method for these languages and compared the results with a baseline characterized by the original English queries and the use of the vector space model. The experimental results showed that our method achieves retrieval performance similar to the baseline, for the languages tested. Table 5 summarizes our results.

⇒ **Insert Table 5 here.**

The main ideas described here (except the generic dictionary) have been applied to an experimental medical search engine⁶ that indexes medical articles published in the last 5 years. The search engine provides interfaces in English, Spanish and Portuguese. Furthermore, the user can navigate in a directory based on the MeSH hierarchy or specify its query directly. In both cases, the information is retrieved in a language-independent way.

⁶<http://www.akwanmed.com.br>

References

- [cid, 1980] (1980). *Classificação Internacional de Doenças - Revisão 9 (Volumes 1 e 2)*. Organização Pan-Americana de Saúde, São Paulo, Brazil. EDUSP - Editora Universidade de São Paulo.
- [mes, 2000] (2000). *MeSH - Tree Structures & Alphabetic List*. National Library of Medicine, Bethesda, MA, USA, 12th edition.
- [uml, 2001] (2001). *UMLS Knowledge Sources*. National Library of Medicine, Bethesda, MA, USA, 12th edition.
- [Aljljayl & Frieder, 2001] Aljljayl, M. & Frieder, O. (2001). Effective arabic-english cross-language information retrieval via machine-readable dictionaries and machine translation. In *Proceedings of the tenth international conference on Information and knowledge management* (pp. 295–302).: ACM Press.
- [Baeza-Yates & Ribeiro-Neto, 1999] Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Harlow, England: Addison-Wesley.
- [Ballesteros & Croft, 1996] Ballesteros, L. & Croft, W. B. (1996). Dictionary methods for cross-lingual information retrieval. In *Proceedings of the 17th International Conference on Database and Expert Systems Applications* (pp. 791–801). Zurich, Switzerland.
- [Cimino, 1995] Cimino, J. J. (1995). Vocabulary and health care information technology: State of the art. *Journal of the American Society for Information Science*, 46(10), 777–782.
- [Dumais et al., 1996] Dumais, S., Letsche, T., Littman, M., & Landauer, T. (1996). Automatic cross-linguistic information retrieval using latent semantic indexing. In *Working Notes of the Workshop on Cross-Linguistic Information Retrieval (ACM SIGIR)* Zurich, Switzerland.
- [Eichmann et al., 1998] Eichmann, D., Ruiz, M., & Srinivasan, P. (1998). Cross-Language information retrieval with the UMLS metathesaurus. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 72–80). Melbourne, Australia.
- [Federico & Bertoldi, 2002] Federico, M. & Bertoldi, N. (2002). Statistical cross-language information retrieval using n-best query translations. In *Proceedings of the 25th International ACM SIGIR Conference on Research*

and Development in Information Retrieval (pp. 167–174). Tampere, Finland.

- [Gao et al., 2001] Gao, J., Nie, J., Xun, E., Zhang, J., M.Zhou, & Huang, C. (2001). Improving query translation for cross-language information retrieval using statistical models. In *Proceedings of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 96–104). New Orleans, Louisiana, USA.
- [Gollins & Sanderson, 2001] Gollins, T. & Sanderson, M. (2001). Improving cross language retrieval with triangulated translation. In *Proceedings of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 90–95). New Orleans, Louisiana, USA.
- [Hersh et al., 1996] Hersh, W., Brown, K., Donohoe, L., Campbell, E., & Horacek, A. (1996). CliniWeb: Managing clinical information on the World Wide Web. *Journal of the American Medical Informatics Association*, 3(4), 273–280.
- [Hersh et al., 1994] Hersh, W., Buckley, C., Leone, T., & Hickam, D. (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 192–200). Dublin, Ireland.
- [Hersh & Donohoe, 1998] Hersh, W. & Donohoe, L. C. (1998). SAPHIRE International: A tool for Cross-Language information retrieval. In *Proceedings of the 1998 AMIA Annual Fall Symposium*, (pp. 673–677). Lake Buena Vista, FL, USA.
- [Lavrenko et al., 2002] Lavrenko, V., Choquette, M., & Croft, W. B. (2002). Cross-lingual relevance models. In *Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 175–182). Tampere, Finland.
- [Lima et al., 1998] Lima, L. R. S., Laender, A. H. F., & Ribeiro-Neto, B. (1998). A hierarchical approach to the automatic categorization of medical documents. In *Proceedings of the 7th International Conference on Information Knowledge Management* (pp. 132–138). Bethesda, MD, USA.
- [Nie et al., 1999] Nie, J. Y., Simard, M., Isabelle, P., & Durand, R. (1999). Cross-Language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proceedings of the 22nd*

International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 74–75). Berkeley, CA, USA.

- [Oard, 1997] Oard, D. W. (1997). Alternative approaches for Cross-Language text retrieval. In *Electronic Working Notes of the AAAI Spring Symposium on Cross-Language Text and Speech Retrieval (1997)*.
- [Oard, 2000] Oard, D. W. (2000). Evaluating interactive cross-language information retrieval: Document selection. In *Cross Language Evaluation Forum (CLEF)* (pp. 57–71). Lisbon, Portugal.
- [Ogden & Davis, 2000] Ogden, W. C. & Davis, M. W. (2000). Improving Cross-Language text retrieval with human interactions. In *Proceedings of the 33rd Hawaii International Conference on System Sciences* Maui, HI, USA.
- [Pearl, 1988] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA, USA: Morgan Kaufmann Publishers, Inc.
- [Rassinoux et al., 1998] Rassinoux, A.-M., Lovis, C., Baud, R. H., & Scherrer, J.-R. (1998). Versatility of a multilingual and bi-directional approach for medical language processing. In *Proceedings of the American Medical Informatics Association 1998 Annual Symposium (AMIA'98)*. (pp. 668–672). Philadelphia, PA, USA.
- [Resnik et al., 2001] Resnik, P., Oard, D. W., & Levow, G. (2001). Improved cross-language retrieval using backoff translation. In *1st International Conference on Human Language Technologies (HLT)* San Diego, CA, USA.
- [Ribeiro-Neto et al., 2001] Ribeiro-Neto, B., Laender, A. H. F., & Lima, L. R. S. (2001). An experimental study in automatically categorizing medical documents. *Journal of the American Society for Information Science and Technology (JASIS)*, 52(5), 391–401.
- [Ribeiro-Neto & Muntz, 1996] Ribeiro-Neto, B. & Muntz, R. (1996). A belief network model for IR. In *Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 253–260). Zurich, Switzerland.
- [Sacaleanu et al., 2003] Sacaleanu, B., Volk, M., & Buitelaar, P. (2003). A cross-language document retrieval system based on semantic annotation. In *Proceedings of EACL 2003 Demo Session* Budapest, Hungary.

- [Salton, 1968] Salton, G. (1968). *Automatic Information Organization and Retrieval*. New York, NY: McGraw-Hill.
- [Salton, 1970] Salton, G. (1970). Automatic processing of foreign language documents. *Journal of the American Society for Information Sciences*, 21(3), 187–194.
- [Salton & Buckley, 1988] Salton, G. & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management*, (24), 513–523.
- [Steffen et al., 2003] Steffen, D., Sacaleanu, B., & Buitelaar, P. (2003). Domain specific sense disambiguation with unsupervised methods. In *GermaNet-Workshops des GLDV-AK Lexikografie* Tbingen, Germany.
- [Whittemore, 1994] Whittemore, G. (1994). The MENELAS English natural language understander: Natural language understanding in the medical domain. In *Proceedings of the 1st World Congress on Computational Medicine, Public Health, and Biotechnology* Austin, TX, USA.
- [Widdows, 2003] Widdows, D. (2003). Unsupervised methods for developing taxonomies by combining syntactic and statistical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 276–283). Edmonton, Canada.
- [Wong & Yao, 1991] Wong, S. & Yao, Y. (1991). A probabilistic inference model for information retrieval. *Information Systems*, 16.
- [Wong & Yao, 1995] Wong, S. & Yao, Y. (1995). On modeling information retrieval with probabilistic inference. *ACM Transactions on Information Systems*, 13(1), 38–68.
- [Yang et al., 1998] Yang, Y., Carbonell, J. G., , Brown, R. D., & Frederking, R. E. (1998). Translingual information retrieval: Learning from bilingual corpora. *Artificial Intelligent Journal*, 103(1–2), 323–345.
- [Zweigenbaum et al., 1995] Zweigenbaum, P., Bachimont, B., Bouaud, J., Charlet, J., & Boisvieux, J. (1995). A Multi-Lingual architecture for building a normalised conceptual representation from medical language. In *Proceedings of the 19th Annual Symposium on Computer Applications in Medical Care* (pp. 357–361). New Orleans, LA, USA.

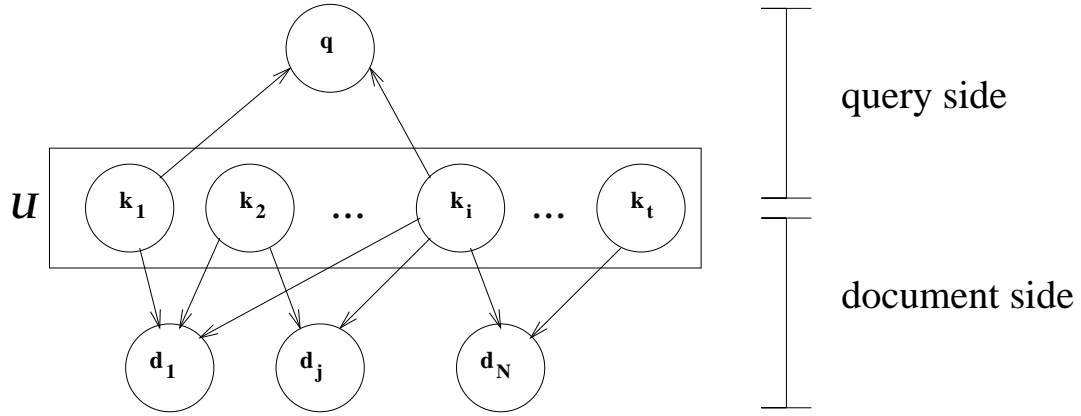


Figure 3: Belief network for a query q given by the keywords k_1 and k_i .

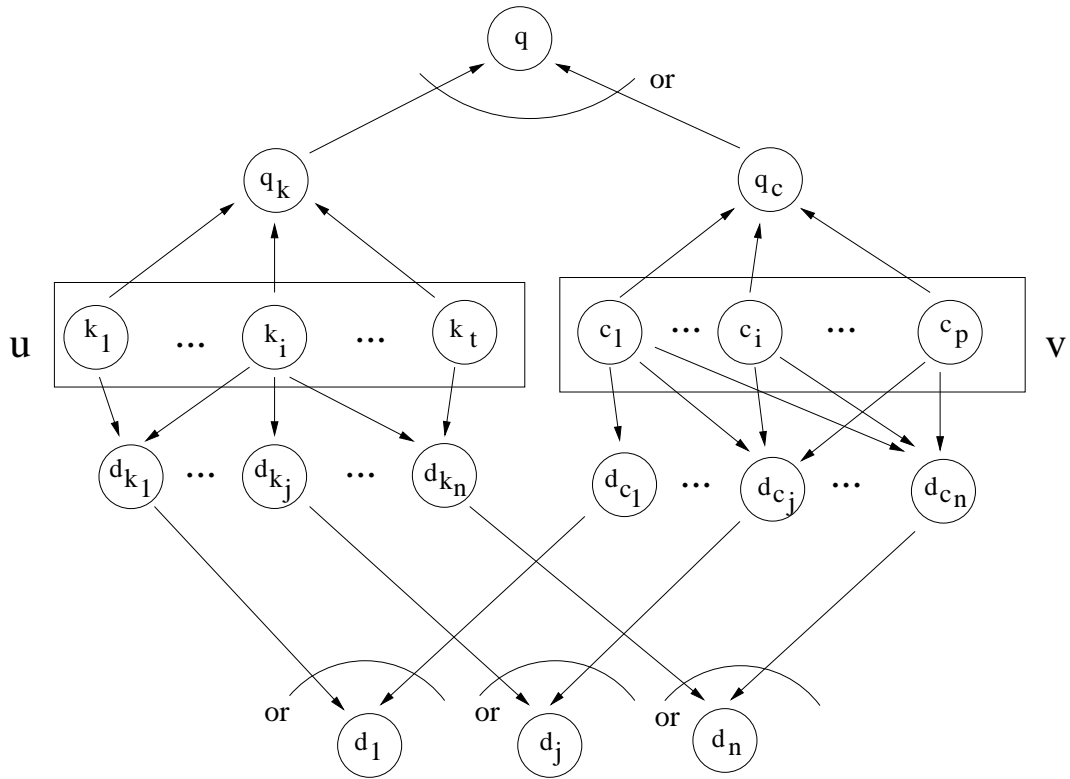


Figure 4: Bayesian network for our ranking fusion model.

Id	8800001
Title	The binding of acetaldehyde to the active site ...
Source	Alcohol Alcohol 8801; 22(2):103-12
Authors	Mauch TJ; Tuma DJ; Sorrell MF.
MeSH Terms	Acetaldehyde/*ME; Buffers; Catalysis; ...
Pub. Type	JOURNAL ARTICLE.
Abstract	Ribonuclease A was reacted with [1-13C,1,2-14C]acetaldehyde and sodium cyanoborohydride in the presence or absence of 0.2 M phosphate. After several hours of incubation at 4 degrees C (pH 7.4) acetaldehyde-RNase adducts were formed, and the extent of their formation was similar stable regardless of the presence of phosphate. Although the total amount of covalent binding was comparable in the absence or presence of phosphate, this active site ligand ...

Figure 5: Example of an OHSUMED document.

Id	24
Query	relationship between prozac and liver disease
Obs	prozac and liver disease
Id	60
Query	treatment of endocarditis with oral antibiotics
Obs	28 yr old male with endocarditis

Figure 6: Examples of OHSUMED queries.

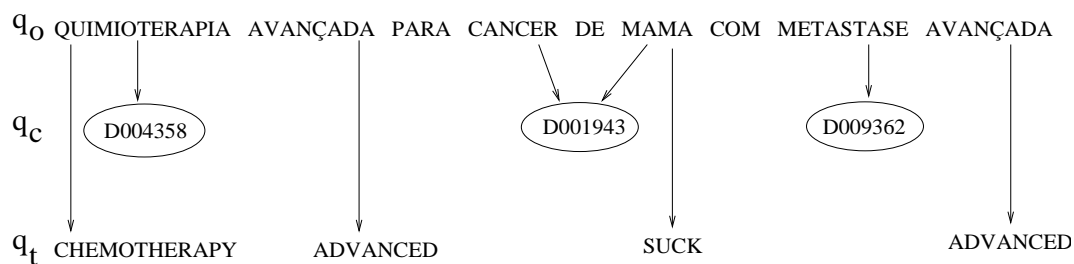


Figure 7: Processing of a query in Portuguese.

	Baseline - $AvgP_{11}$	
	cos	nfx.bfx
q_b	0,2433 (100,0%)	0,2333 (95,8%)
q_c	0,2185 (89,8%)	0,2298 (94,4%)
q_{bc}	0,1854 (76,2%)	0,2645 (108,7%)

Table 1: Baseline evaluation.

$AvgP_{11}$			
	Portuguese	Spanish	French
q_o	0,0542 (22,3%)	0,0656 (26,9%)	0,0673 (27,6%)
q_c	0,2237 (91,9%)	0,2403 (98,7%)	0,2031 (83,4%)
q_t	0,0803 (33,0%)	0,0803 (33,0%)	0,0703 (28,8%)
q_{oc}	0,2267 (93,1%)	0,2362 (97,0%)	0,2236 (91,9%)
q_{ct}	0,2384 (97,9%)	0,2346 (96,4%)	0,2169 (89,1%)
q_{ot}	0,1144 (47,0%)	0,1264 (51,9%)	0,1281 (52,6%)
q_{oct}	0,2422 (99,5%)	0,2480 (101,9%)	0,2300 (94,5%)
Baseline = 0,2433 (100%)			

Table 2: $AvgP_{11}$ evaluation of our method and comparison with baseline.

Recall	Precision			
	Baseline q_b (cos)	Portuguese q_{oct}	Spanish q_{oct}	French q_{oct}
0,00	0,4441 (100%)	0,4512 (101,5%)	0,4758 (107,1%)	0,4124 (92,8%)
0,10	0,3846 (100%)	0,4161 (108,1%)	0,4225 (109,8%)	0,3774 (98,1%)
0,20	0,3364 (100%)	0,3494 (103,8%)	0,3656 (108,6%)	0,3295 (97,9%)
0,30	0,3052 (100%)	0,2915 (95,5%)	0,3053 (100,0%)	0,2970 (97,3%)
0,40	0,2790 (100%)	0,2653 (95,0%)	0,2729 (97,8%)	0,2653 (95,0%)
0,50	0,2205 (100%)	0,2189 (99,2%)	0,2211 (100,2%)	0,2084 (94,5%)
0,60	0,2041 (100%)	0,2030 (99,4%)	0,2040 (99,9%)	0,1900 (93,0%)
0,70	0,1729 (100%)	0,1608 (93,0%)	0,1646 (95,1%)	0,1544 (89,3%)
0,80	0,1454 (100%)	0,1309 (90,0%)	0,1270 (87,3%)	0,1276 (87,7%)
0,90	0,1247 (100%)	0,1078 (86,4%)	0,1041 (83,4%)	0,1088 (87,2%)
1,00	0,0597 (100%)	0,0688 (115,2%)	0,0654 (109,5%)	0,0597 (100,0%)
$AvgP_{11}$	0,2433 (100%)	0,2422 (99,5%)	0,2480 (101,9%)	0,2300 (94,5%)

Table 3: Recall x Precision evaluation.

Top_nP	Precision			
	Baseline	Portuguese	Spanish	French
	q_b (cos)	q_{oct}	q_{oct}	q_{oct}
Top_5P	0,2257 (100%)	0,2475 (109,6%)	0,2514 (111,3%)	0,2099 (92,9%)
$Top_{10}P$	0,1970 (100%)	0,2148 (109,0%)	0,2207 (120,3%)	0,2019 (102,4%)
$Top_{20}P$	0,1861 (100%)	0,1767 (94,9%)	0,1821 (97,8%)	0,1712 (91,9%)
$Top_{30}P$	0,1702 (100%)	0,1607 (94,4%)	0,1630 (95,7%)	0,1544 (90,7%)
$Top_{50}P$	0,1491 (100%)	0,1370 (91,8%)	0,1413 (94,7%)	0,1356 (90,9%)

Table 4: Top_nP evaluation.

	Spanish	Portuguese	French
$AvgP_{11}$	101%	99%	94%
10% recall	109%	108%	98%
$Top_{10}P$	120%	109%	102%

Table 5: Average Precision figures at 11-point recall, 10% recall, and top 10 documents.