Computing and Information Sciences

Recent Trends

Editor
J.C. MISRA

Associate Editors
A. GOSWAMI
PAWAN KUMAR

Narosa

Computing and Information Sciences: Recent Trends
Edited by J.C. Misra

7. Intelligent Information Retrieval Systems

M.T. Martín-Valdivia, L.A. Ureña-López, F. Martínez-Santiago and M. García-Vega

Departamento de Informática. University of Jaén. Campus Las Lagunillas, s/n E-23071. Jaén, Spain

1. Introduction

Information retrieval (IR) systems search in a document collection in natural language, with the objective of retrieving exactly the subset of documents that answers an user's query. Unlike database systems, which require highly structured data, IR systems work with unstructured text written in natural language. In terms of their differences with expert systems, IR systems do not try to deduce nor generate specific answers, but return documents whose content is similar to the user's query. Although IR systems have existed since the 1960s, popular World Wide Web search engines (Google, Alta Vista, InfoSeek, etc.) are current examples of the complexity of the IR task.

This task is more relevent when one considers the proliferation of texts available on the Internet (it is estimated that in 10 years the amount of information written in history will have doubled) [19]. Thus it is more and more necessary to be able the access information based on the semantic content of the next.

A large number of IR systems use statistical and information theory techniques, although these techniques have recently been combined with linguistic resources and Natural Language Processing (NLP) techniques.

There are many definitions of IR systems. One of the best has been formulated by Salton and McGill [34] where they show that IR deals with representation, storage, organisation and access to information items. The entries to the system are the documents of the database and the user query. The result is a subset of the documents classified as relevant, that is to say, as suitable to the user's needs. Usually, a numerical value is assigned to each relevant document. The higher the number, the more relevant the document.

This chapter is structured as follows: first, a description of the main models and indexation techniques used in IR. Secondly, we comment on the linguistic resources related to IR. Thirdly, we describe IR applied to texts written in different languages. Finally, we present the main forums dedicated to the evaluation of different IR systems.

2. Information Retrieval

2.1 Introduction

Information retrieval (IR) is the selection of the subset of suitable documents for

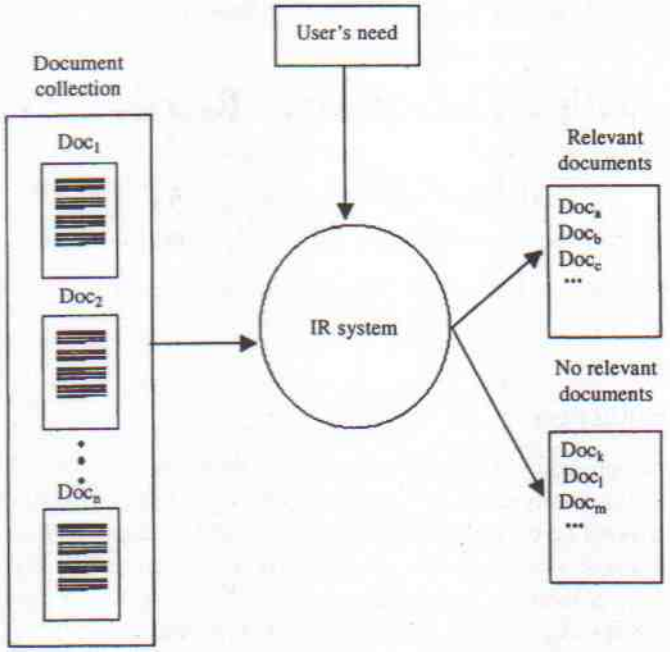

Fig. 1. Information retrieval systems.

the needs of a user between a wider set in a document database [34]. The user's information need is usually represented through a natural language query. The system typically selects those documents or text that contains the query terms.

The IR process includes several tasks. One of them is to obtain a suitable representation of the documents, which allows their efficient storage and comparison with the user's query (indexing process). Another tass is the user query processing. Any query must be represented in a more or less sophisticated query language. The last task consists of the comparison of the query and the documents to obtain the document list that satisfies the query [7]. Although the mechanisms developed up to now for access to that, information are constantly improving due to new interfaces and facilities, the information retrieval of largest electronic resources is based, in general, on the closeness of keywords or fixed indexation.

2.2 Models

An IR model must have a determined representation system, both of the documents that it aims to recover, and of query formulated. In addition, it must have determined the measure that calculates the relevance of a document with regard to a given query. The relevant documents are shown according to this measure [1]. Formally, we determine the IR model as follows:

$$\text{IRM} = [D, Q, F, R(q_i, d_j)]$$

where D and Q are the documents and queries respectively, F is the documents and queries domain, as well as their relations and R the similarity function that

associates to each document a real number according to a given query, determining the order of documents for that query.

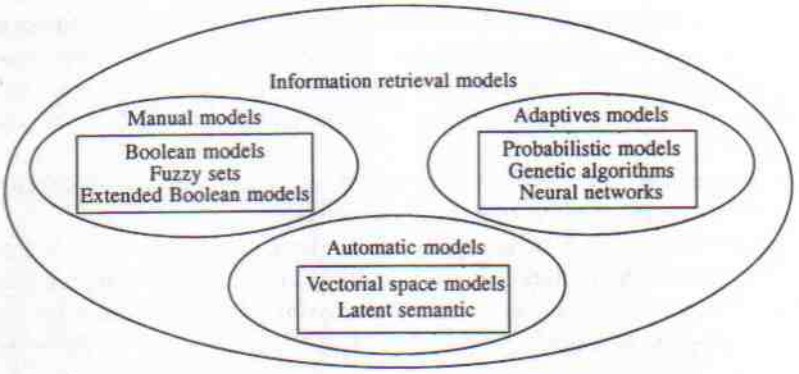

Fig. 2. Information retrieval models.

The different interpretations that are given to this determination appear in several conceptual models. We can classify them into manual systems, where the queries must be made by an expert user (Boolean Model), automatic systems, where it is only necessary to use a list of simple words in a query (Vectorial Space Model), and, finally, adaptive systems, which allow better results using training documents (Probabilistic Model).

2.2.1 *Boolean Model*

This model has dominated the IR systems for more than 30 years [25]. The majority of IR commercial systems are built with an engine based on it, undoubtedly because the search engines with this approach are fast, use few resources and are easy to implement. However, it has large disadvantages, such as difficulty of management, poor results obtained recall and precision, and the lack of sorting function that orders the returned documents.

This model represents, on the one hand, the documents as a word set or index terms, and, on the other hand, the query as a Boolean expression, with the operators AND, OR and NOT. So, for any query, the relevant documents contain the terms that are included in the corresponding expression. The words simply appear or not, with no reference to their frequency in the documents.

To calculate whether a document satisfies a Boolean expression, and therefore, if it is relevent, is enough to calculate the disjunctive normal form of the expression and check some of these elements.

Formally, let $\mathbf{W} = \{w_1, w_2, \ldots, w_n\}$ be the set of index terms, and let $g(i, \vec{\mathbf{d}})$ be the function that returns TRUE if w_i is in $\vec{\mathrm{d}}$. We say that a $\vec{\mathbf{d}}_j$ document is relevant for the $\vec{\mathbf{q}}$ query, where $\vec{\mathbf{q}}_{\mathrm{dnf}}$ is the disjunctive normal form, and q_c, any of its components, if $\forall\, w_i, g(i, \vec{\mathbf{d}}) = g(i, \vec{\mathbf{q}}_c)$.

2.2.2 *Vectorial Space Model*

The use of Vectorial Space Model (VSM) has a 3-stage process: the indexation

of all the terms of the documents, the assignation of weight to these terms and the order of the documents with respect to the query.

Perhaps, the main advantage of this model is that the queries are a sequence of terms and, therefore, it is easy to use. In addition, these 3 stages are automatic. This model is based on the interpretation of each document of the collection as a n-dimensional vector, where n is the number of different terms that are in the document collection and calculated with the classic weights tf-idf, based on the frequency of the term [34].

In another way, the query is a collection of terms, so it could be dealt with in the same way, converting all the text into vectors.

To determine which documents are relevant for a given query it is enough to check which are the vectors that are near the vector that represents the query. Hence it is only necessary to calculate the cosine of the angle that form the vectors using the formula:

$$\text{similarity}\,(\vec{d_j} \cdot \vec{q}) = \frac{\vec{d_j} \cdot \vec{q}}{\vec{d_j} \times |\vec{q}|} = \frac{\sum_{i=1}^{n} wd_{ji} \times wq_i}{\sqrt{\sum_{i=1}^{n} wd_{ji}^2 \times \sum_{i=1}^{n} wq_i^2}}$$

where wd_{ji} is the weight of the ith them of the document, wq_i the weight of the ith term of the query and n the number of different terms in the collection of documents.

This model is notable for its simplicity and can directly include relevance feedback, which makes it a very powerful model.

For example, supposing the weights of the terms equal to the frequency:

q: "the white horse"
d_1: "the white horse rides"
d_2: "the white colour"

$$\text{sim}(\vec{d}_1, \vec{q}) = \frac{1 \times 1 + 1 \times 1 + 1 \times 1 + 0 \times 1 + 0 \times 0}{\sqrt{4 \times 3}} = 0.87$$

$$\text{sim}(\vec{d}_1, \vec{q}) = \frac{1 \times 1 + 1 \times 1 + 0 \times 1 + 0 \times 0 + 0 \times 0}{\sqrt{4 \times 3}} = 0.67$$

2.2.3 *Probabilistic Model*

The Probabilistic Model (PM also uses the concept of vectorial space and the similarity between documents. Further more, it adds the relations between the different terms, so, while in the VSM the different axes of coordinates were orthogonal, in the PM, the dependence between them is taken into accounts.

The main idea of this model is found in the way that, for any query, the terms of the relevant documents that have been found in a previous search must have greater weight than those which did not appear.

So, the collection of documents could be divided in two groups: relevant and not relevant documents, for any given query, thereby finding the ideal group of documents. As at first, the properties of this ideal group are unknown, the user

can improve the first description, recycling and reiterating the process, until all the relevant documents are found.

Therefore we must calculate, for a given q query, the probability that a d_j document is relevant. If we define the relevant set of documents as R_q, we must calculate the similarity according to:

$$\text{similarity}\,(\mathbf{d_j}, \mathbf{q}) = O(R_q \mid d_j) = \frac{P(R_q \mid d_j)}{P(\overline{R_q} \mid d_j)}$$

and applying the Bayes theorem, we obtain

$$\text{similarity}\,(\mathrm{d_j}, \mathrm{q}) = \frac{P(R_q)}{P(\overline{R_q})} \times \frac{P(d_j \mid R_q)}{P(d_j \mid \overline{R_q})}$$

The typical weight of any k_i term is calculated as $P(k_i \mid R_q) = 0.5$ and $P(k_i \mid \overline{R}_q) = \frac{n_i}{N}$, where n_i is the document number where k_i terms appears and n is the total number of documents.

The probabilistic model is mainly applied in relevance feedback, because it performs better if a good training is used. However, it requires a high effort process.

2.3 Techniques of Indexation

2.3.1 Weight Terms

Different techniques have been developed to calculate of the weight the terms which must internally represent the documents. The majority of these methods are based on the concept of resolution power of a term[1], using as a measure its adequacy as an indexation term [34, 33] and to reduce the document vector dimension. [26] establishes a relationship between the discrimination degree or resolution power and its frequency in the document. So, the words with higher resolution power have a medium frequency of appearance. The justification for the elimination of infrequent terms is based on an observation made by [44] and known as "Zipf's law"[2], about the frequency of words in a corpus of texts, which establishes that by ordering the words in a text (or collection of text) by their frequency of one use, the product of that frequency by its position in the order is constant. That relation is graphically shown in Fig. 3.

One way to get an assignation of weights near to the definition of resolution power is proposed by [35]. This is based on the assignation of weights to the terms that appear in the documents by the following expressions:

$$wd_{ij} = tf_{ji} \cdot idf_i$$

[1]It gives a base for the indexation methods based on the frequency of appearance of the terms.

[2]That is currently not a law, only an empire phenomenon.

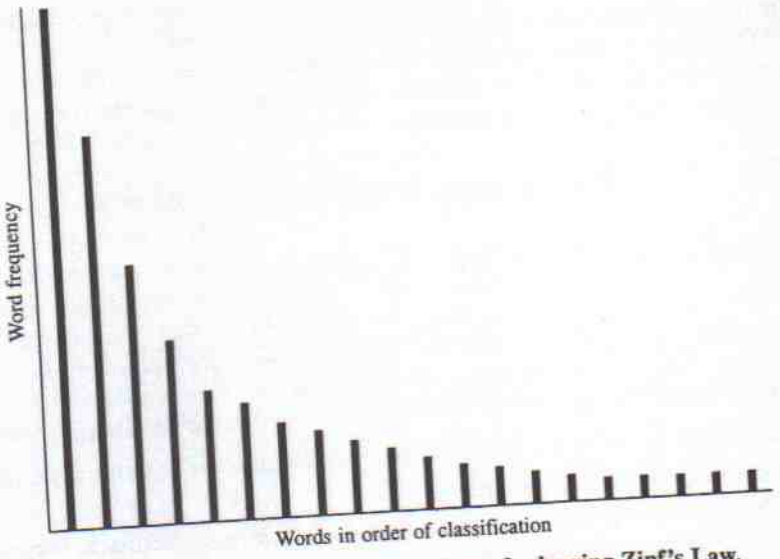

Fig. 3. Histogram of the frequency of words showing Zipf's Law.

$$idf_i = \log_2\left(\frac{n}{df_i}\right)$$

where wd_{ji} is the value of the weight assigned to the i term in the j document, tf_{ji} is the frequency of appearance in the document j of the i term, df_i, is the document frequency of i term, understood as the number of documents where that term appears, and idf is the inverse of the document frequency for the i term. In the expression at calculates this last value, n is the number of documnets of the collection.

2.3.2 *Stoplist*

A we have seem in the previous section, there are a set of terms, with common use, that do not have resolution power, and therefore are not good candidates to be part of an index of terms.

To eliminate the terms that do no contribute to resolution power and so are not useful to the index process, stoplists [11] have been used since the beginning of IR to increase the effectiveness of the IR process. Some examples of this words in English, with little capacity of discrimination and little representation as indexation terms are: "the", "of", "and", "to", etc.

These empty words are often in the documents, as is shown by the fact that the ten words most frequently used in English, could form between 20% and 30% of the terms in a document [14].

These lists are obtained in specifically oriented studies, from a corpus of texts representative of the considered language. For example, we can found in [33] a stoplist of 250 terms, and in [11] one of 425, obtained from BROWN CORPUS [9].

a	all	and	are	became
a's	allow	another	aren't	because
able	allows	any	around	become
about	almost	anybody	as	becomes
above	alone	anyhow	aside	becoming
according	along	anyone	ask	been
accordingly	already	anything	asking	before
across	also	anyway	associated	beforehand
actually	although	anyways	at	behind
after	always	anywhere	available	being
afterwards	am	apart	away	believe
again	among	appear	awfully	below
against	amongst	appreciate	b	beside
ain't	an	appropriate	be	besides

Fig. 4. First terms of the stoplist used in Smart[3].

2.3.3 Stemming

The extraction of roots (stemming) is a technique used to improve the performance of the IR systems, because it increases the effectiveness of the IR process and reduces the size of the index archieves[4]. The goal of stemming algorithms is to obtain a single term from several words with the same meaning that essentially differs in its morphology [12, 23]. For example, we could consider obtaining the term INFORM, from "information", "inform", "informer" and "informed".

The result of the algorithm must be a single canonical form for the different morphologic variations of a word which need not necessary be the linguistic root. In the bibliography different kinds of stemming algorithms can be found.[5][12].

Two kinds of errors can be introduced in this kind of algorithm:

- Understemming, which results from obtaining different canonical forms for words which should give a single root due to having the same meaning, for example, for "*hope*" and "*hopping*" HOP and HOPP are obtained.
- Overstemming, which results from obtaining the same canonical form for words that should have different forms, as they differ in meaning but not in morphology significant, for example obtaining CAPIT from "*capital*", "*capitulate*" and "*capitol*".

To avoid mistakes in the IR process, some algorithms to minimise overstemming and understemming are interesting. Less sophisticated algorithms that tends to produce overstemming and avoid understemming (the number of index terms are decreased) could be suitable for systems in which time and space are a priority, but such algorithms produce a loss of the process effectiveness [12].

[3]Smart is an IR system originally developed by Salton at Cornell University. Could be obtained by ftp (ftp://ftp.cs.cornell.edu/pub/smart).

[4]The reduction is about 50% [12].

[5]One of the more widely known is from [32]. It is characterised by its small size, and by the simplicity of the set of rules to remove suffixes. This simple algorithm satisfactorily substitutes a morphological analysis for the process of text retrieval.

2.3.4 Feedback

The concept of feedback can also be used in text retrieval systems [17, 18] to improve the efficiency of IR systems. This implies allowing the user to formulate a query by the selection of a subset of documents of the database that he specifies to the system to be suitable to his needs. This specification could be summed up as "find more documents similar to this (or these)".

3. Linguistic Resources

Many studies attempt to integrate linguistics resources in a model that allows their use and application to IR tasks [39]. The basic idea consists fo including as much information as possible from linguistic resources or through typical techniques of NLP, with the purpose of obtaining more effective systems.

New approaches have been developed based on linguistic resources as sources of external knowledge. In fact, both disambiguation systems and categorisation systems have improved in effectiveness using information given by these resources.

Although researchers have attempted to apply this idea since the appearance of the first IR systems, it was the 1990s which saw a revival in the interest in using linguistic resources to improve the IR task. This happened because of the efforts of many organisations and associations (Association for Computational Linguistics Data Collection Initiative -ACL/DCI-, European Corpus Initiative -ECI-, British National Corpus -BCN-, Linguistic Data Consortium -LDC-, etc.) to make available a large amount of information of linguistic kind as text corpora (a collection of representative texts), Machine Readable Dictionaries (already existing and published in paper but available in electronic form), thesaurus (data structure that groups synonymous terms), etc. with the purpose of improving the processing and evaluation of IR systems.

Here we describe two of the most widely used linguistic resources: text corpora and lexical databases.

3.1 Corpora

A linguistic corpus is a collection of texts representative of a language, of a dialect, or of a language subset which are used for linguistic analysis [13].

Although corpora are available since several decades ago, it is in recent years when their use has became popular and extended due to the good results obtained when corpora are used in IR systems, and the possibility of their computational use in research tasks, thanks to their availability in electronic form.

Among the features of a corpus, we can highlight the size, diversity and balance of the examples included and the precision and variety of the lexical and linguistic information levels that it represents.

The corpora can be classified by many criteria. First, the material that incorporates can be distinguished by two large groups, text and oral, depending on whether they are recompilations of written texts or transcriptions of oral language.

Secondly, depending on the purpose, the text corpora can be considered as general purpose or specific purpose. General purpose corpora are constructions

of language for general applications, whereas specific purpose corpora are centred on a specific need.

Thirdly, two further subcategories can be established: general language corpora and sub-language corpora, depending on whether they combine diverse kinds of text or limit their content to a specific kind of text.

Finally, the corpus may be linguistically annotated or non-annotated. This is the most widely-used classification. A non-annotated corpus has available only a collection of texts without any additional information, while an annotated corpus provides aditional information to the text in form of notes or annotations. Annotated corpora have been a great advance in the use of NLP resources in IR tasks, because they contain a high linguistic value, showing annotations of different language levels. The annotation could be made in many linguistic levels (grammatical, semantic, etc.). The first annotated corpora were generated manually, but more recently annotations have been made semi- or totally automatic (using taggers and parsers). Manual corpora are higher quality and have fewer errors, but imply using a larger amount of human resources.

Among non-annotated corpora, a prime example is the BROWN CORPUS [14]. Among annotated corpora, SEMCOR (Semantic Concordance) [28] is considered one of the most useful by researchers due to its availability and high coverage. It is a subset of the BROWN CORPUS that is annotated on a syntacxtic and semantic level. The labelling has been manually done with the meanings of the words defined in the lexical database WORDNET. SEMCOR 1.5 has 500 passages of 2,000 words each, taken from contemporary publications of documents. It was designed as a heterogeneous and balanced collection of texts, with different styles and literary genres, dealing with political, scientific, literary, sporting, musical and other topics.

Another annotated corpus recently made available is the Reuters document collection [24], containing more than 20, 000 news items of an economic character featured on the Reuters news Channel during year 1987.

As we have mentioned, multilingual IR systems have recently become more important, fundamentally due to the massive use of Internet by the monolingual users of different countries who require information, regardless of the language in which that information is written. Multilingual resources, such as parallel corpora, have special interest in these cases.

A parallel corpus refers to a collection or set of texts where each is the translation from the original to other languages. The simplest example uses two languages (bilingual corpus), and consists of two corpora, each one in one language, where one is the exact translation of the other. An example is the Canadian Hansard, which consists of 500 million transcriptions in French and English of the proceedings of the Canadian Parliament. Among examples of parallel corpora in many languages can be seen the polyglot Bible [8], in which St. Luke's gospel is written in 13 different languages.

The World Wide Web is an inexhaustive source of linguistic resources. There currently exists a growing interest in corpus generation extracted from the Web [16]. An initiative along these lines is describe in [27].

3.2 Lexical Database

A lexical database is a system whose main objective is to store information about a term set in one or more languages. Lexical databases differ in language, number of languages and the kind of information for each term.

One of the most widely-used lexical databases by many applications and researchers is WORDNET [10, 28, 29]. WORDNET[6] is a system developed by the Cognitive Science Laboratory at the University of Princeton, which is inspired by psycholinguistic theories. It is a system which contains lexical information extracted semi-automatically from dictionaries and which is available on Internet.

In WORDNET, the basic element which allows the representation of concepts as synonymous sets is the synset, so as well as including information about each term, information concerning the different relationships between words and synsets is stored, such as hyponymy (generalisation relationship, is-a), meronymy (inverse of the has—a relationship) and antonymy, as well as synonymy which is implicit in synset definitions.

WORDNET also stores syntactic information; each term is designated by category: noun, verb, adjactive or adverb.

WORDNET only includes information for English terms. However, there exist other lexical databases used in multilingual IR systems which includes more than one language. For example, ACQUILEX [40] uses information extraction techniques from more of 10 dictionaries in 4 languages of the European Community; EDR (Electronic Dictionary Research) [43], is a project developed in Japan that includes English and Japanese dictionaries, as well as bilingual English–Japanese and Japanese–English dictionaries; EUROWORDNET[7] [42] is a project that includes many languages of the European Community (Dutch, Italian, Spanish, German, French, Czechoslovak and Estonian) inspired by the lexical database WORDNET. However, EUROWORDNET is a multilingual database rather than a semantic database set for each language, where each language is connected with all the others, over an Inter-Lingual-Index.

4. Cross-Language Information Retrieval

4.1 What is CLIR?

Since the second half of the 1990s, Cross-Language Information Retrieval (CLIR) has become more prominent in the IR community and it is today a discipline which receives as great an interest as traditional IR. A CLIR system is basically an IR system capable of operating over a cross-lingual documents collection. That is, if a user consults a CLIR system, all relevant documents in the collection are retrieved, independently of the language used in the query and the documents. So the result of one of this systems will frequently be a heterogeneous list of documents written in English, Spanish, French, German, etc. and ordered according to the score given to each document for a given query.

---

[6] http://www.cogsci.princeton.edu/-wn

[7] http://www.hum.uva.nl/-ewn/

The growing interest in multilingual systems is fundamentally due to two reasons: first, the popularity of Internet has made the Net an enormous multilingual collection of documents; and second, in the global society, multinational organisations generate large amounts of documents, usually written in the native languages of different regions where the organisation is present. The typical user of the multilingual system will be someone with notions of languages present in the collection of multilingual documents, but without enough ability to express their need for information by means of a specific query in each language.

4.2 CLIR Problems

If the CLIR task is about selecting relevant documents for determined information needed by the user, in a multilingual process it is also necessary to overcome the linguistic barrier that appears between the query language and the different languages present in the collection being consulted [31]. So, any serious attempts at developing a CLIR system which is able to obtain similar results to those of a monolingual system must take into account the following points, which we will refer to as the 3 CLIR problems:

1. Translation of the query and the documents.
2. Usually, we obtain more than one translation for each query. When the translations have been carried, which should be chosen?
3. How can we obtain a single list of relevant documents, independently of the language used in each document?

We can note that how to translate and how to reject less specific translations are typical problems in Machine Translation (MT). However, a CLIR system is less demanding than a MT system in terms of the quality of the translations. It has been empirically shown that while an MT system obtains its best results taking as a unit the sentence, the IR system seems to produce better results taking the word as the base unit and disregarding links between words. Therefore CLIR systems centre all their efforts on obtaining the most accurate set of possible translations for each term, in order to solve the 3 CLIR problems.

Usually, CLIR systems are classified according to translation techniques. About this classification deals next sections.

4.3 Translation of Queries and Documents

The textual elements with which any IR system works, are queries and documents. In a multilingual environment, we must first decide what textual elements to translate, and then how to translate them. We can choose to translate the query, the documents or a mixed approach [31].

4.3.1 Translation of the Query

If we choose to translate only the queries, it is necessary to divide the collection of multilingual documents into a set of monolingual collections. All the documents expressed in a same language will form each monolingual collection. When all the collections are indexed, the selection of documents in a given query will require the translation of the query into the same number of languages there are

monolingual collections. Finally, each translation will be checked with its monolingual collection.

With that approach we don't obtain, in a given query, a single list of relevant documents, but rather a list for each monolingual collection. Bearing in mind that the relevance of each document is obtained in relation to the collection of monolingual documents to which it belongs, and not in relation to the original multilingual collection, How can we combine the lists of documents obtained for each language in a single list? This is the third CLIR problem and it is an open problem. Although various strategies have been experimented with, from applying a simple round-robin like algorithm to normalising the score obtained for each document, the loss of accuracy between 20% and 40%, with respect to IR systems which work with single large collection [41], [5].

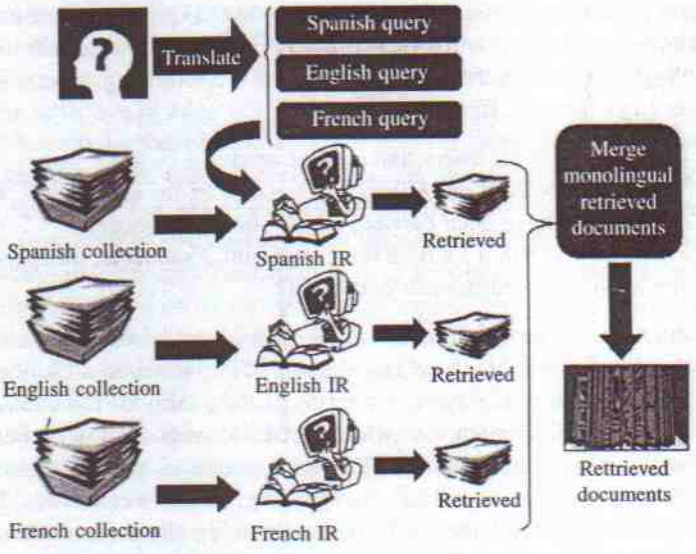

Fig. 5. Query translation-based CLIR model.

4.3.2 Translation of the Documents

An alternative strategy is to establish a "pivot language" in which both the documents and the query must be expressed. That pivot language is usually English, because this language has more linguistic resources. That approach requires the translation of the whole document collection to the pivot language, but simplifies the translation of the query, because it only requires a single translation to the pivot language. We can note that by translating the whole collection to a single language, a single list of relevant documents is the immediate result, implicitly solving the last of the 3 CLR problems. Two of these systems are described in [3] and in [15].

4.3.3 Mixed Approach

This approach is perhaps the most original of the 3, because it ignores the multilingual character of the collection. As in IR and CLIR systems, each document

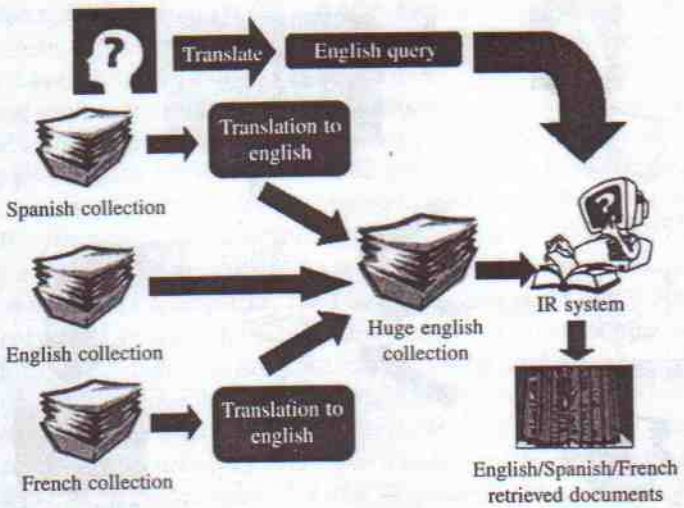

Fig. 6. Documents translation-based CLR model.

is pre-processed, removing frequent terms and obtaining the roots of others. This process is evidently dependent on the language of the document. But from that stage onwards, the multilingual collection of documents is handled by the CLIR system as a monolingual collection, generating a single index of documents in which are mixed terms in different languages, without any translation. The subsequent problem is therefore how to compare a monolingual query with an index of documents mixed in different languages. The answer is to obtain a query in which there are also mixed terms in different languages. That is, the query must be translated into each of the languages of the collection of documents, but unlike the approach based on the translation of the query, we do not generate a query for each translation. Instead, all the translations are mixed to form a single query, as if all the documents were mixed in a single index. This query will be the one which is compared with the collection of documents.

As in the approach based on the translation of the documents, the third CLIR problem is eliminated, because when a single index of documents is generated, a single list of documents is returned by the system for each query. A good example of this approach is the system depicted in [6].

4.3.4 Advantages and Disadvantages of Each Approach

No one approach is clearly better than the others. The choice of one or another will be usually determined by the linguistic resources available and by the specify needs of each moment.

The approach based on the translation of documents has the advantage that a translation of the whole document will in theory be of a higher quality than a translation of the query, as there is more contextual information available, given that queries are usually formed by no more than 2 or 3 words. However, the time needed to translate documents is considerable, even though this task can be carried out off-line, and will therefore not decrease with the performance of the system.

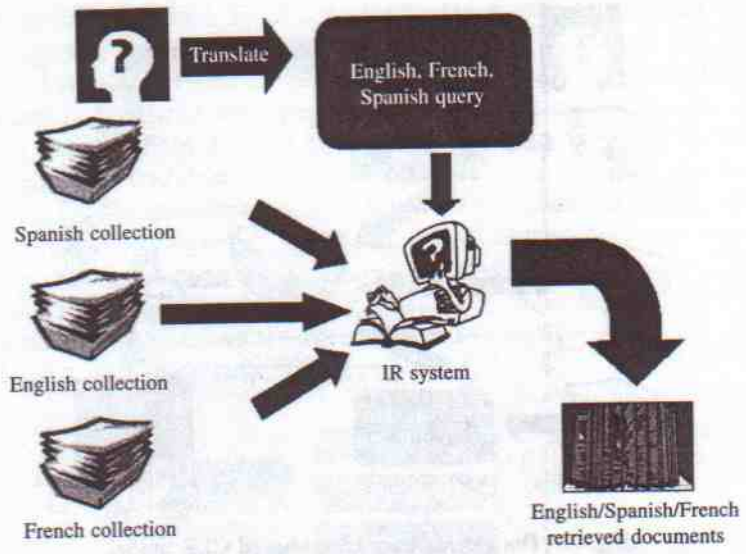

Fig. 7. Mixed CLIR model.

A more serious problem of systems based on a single index of documents (translation of documents and mixed) is the difficulty of scaling: a new document implies re-indexing the whole collection to obtain a single index. This, along with the time necessary for the complete translation of the multilingual collection of documents, means that in general it is preferable to experiment using the approach based on the translation of the query, which is more dynamic and therefore more appropriate to experimentation than the other two.

4.3.5 Translation of Textual Units

When we decide to translate the documents, the query or both, is necessary to choose a resource or tool to start the translation. To value each of the options, we must consider the following aspects [2]:

1. The translation of a term has problems of ambiguity. If T is a term in the original language, the correct translation W will depend on the sense of T.
2. The coverage of the linguistic resource used in the translation. That is, the percentage of terms that are still to be translated, simply because they are not present in the resource.
3. The translation of so-called multiwords (adjacent terms, usually nouns, whose meaning is not the sum total of the meanings of their component terms) is especially complex, because it is common for the translation of a multiword not to be the same as the translations of its component terms. This problem reflects on the accuracy obtained [20].

An immediate approach would be to use an MT system currently on the market, for example, Systran, which is widely used by CLIR community. Although this approach is usually suitable and gives good results, cannot always be applied, because an MT system is a difficult resource to obtain when we intend to work

English/Spanish/French
retrieved documents

ingle index of documents
f scaling: a new document
single index. This, along
the multilingual collection
to experiment using the
ch is more dynamic and
e other two.

/ or both, is necessary to
ıe each of the options, we

uity. If T is a term in the
d on the sense of T.
e translation. That is, the
ɔly because they are not

ıt terms, usually nouns,
their component terms)
ınslation of a multiword
nt terms. This problem

ystem currently on the
IR community. Although
ınnot always be applied,
/hen we intend to work

with little-used languages. In addition, the quality of the translation depends greatly on the two languages considered. Finally, an MT system gives a single translation per term, which is not only unnecessary in the CLIR task, but also is not always convenient. For example, the French term "traitement" could be translated into English as "salary" or "treatment". A MT system must choose one of them, while a CLIR system could consider both. If the original query refers to "waste treatment", then the suitable translation will be "treatment". So, the CLIR system will generate some noise, but this is always preferable to choosing a perhaps incorrect single term ("salary"), which would be the case of the MT system. This example explains the second of the 3 CLIR problems: to what extent should we prune the available translations for each term? It is necessary to search for the balance between the possible noise that we introduce in the translation and the conservation of the original sense of the query.

A no-less attractive approach is to make the translation word by word, using a Machine Readable Dictionary (MRD) or a lexical multilingual database, such as EuroWordNet [42]. An example of a CLIR system based on MRD is presented in [20].

Systems based on MRD have the advantage over an MT system of being a less scarce resource, often with more coverage. However, with this approach, the second of the 3 problems is harder to resolve, because, in general, an MRD will offer a full set of translations for each term. How can we disregard some or all of the inadequate translations? Two options are possible:

1. Given a term T, that could be translated by $S_1, \ldots S_n$, calculate the probability that each S_i is the translation of T.
2. Choose the translation or translations in function of the context in which T appears.

Dictionaries that have this information are very scarce, making it necessary to use resources that complement or substitute them. One of the most valuable resources for this task are parallel corpora. These are multilingual corpora whose documents are exact translations of each other. It is equivalent to having a monolingual corpus, with its translation to another languages. It is also very useful to align them to sentence level, such that for any sentence it is possible know the exact translation in the other parallel documents. We can therefore know the frequency with which a term is a translation of another and also the contexts that often accompany it. The problem of parallel corpora is that they are a difficult resource to find, above all bearing in mind that in order to be really useful, they must contain at least several thousand documents for each considered language. A successful system that uses only parallel corpora in translation is described in [30].

Easier to obtain and not without interest as a resource in translation, are comparable corpora. The restriction is not now to have a corpus translated into several languages: it is enough for the various monolingual corpora to have a common topic, but they do not need to be translations of each other. From a comparable corpus it is possible to generate a thesaurus of similarities, which allows for measurements of similarity between two terms, according to the

context in which they appear [37]. This, in a multilingual environment, could be read thus: if a term *T* often appears in documents that deals with given subject, and a term *S*, in another language, is usually found in documents with the same topic, then perhaps *S* could be a good translation of *T*. This approach has been used with success in [4], and is interesting when resources such as MTs, MDRs or parallel corpora, which are more difficult to obtain than a comparable corpus, are not available.

The tendency of current CLIR systems is integration: to combine the available resources in each moment to obtain the most accurate translation possible. Thus, there are systems which mix MRDs and MTs [36], parallel corpora and MRDs etc.

4.3.6 Summary

In summary, it seems clear that the tendency of recent years is the generation of huge collections of documents written in a heterogeneous set of languages. Therefore, it is necessary overcome the linguistic barrier that arises when we search for information about that collection. To achieve this, there are CLIR systems that translate queries into the necessary languages, or create a monolingual document collection by the translation of the original multilingual collection, or have a mixed approach, translating the queries, but supporting an single index of multilingual documents.

Although the option of translating only the queries seems to be currently predominant, it makes it more difficult to obtain of a single list of relevant documents, because, in general, we obtain as many lists as there are languages in the collection. Translating documents presents problems of scalability and it is cumbersome to translate the whole collection, especially in an experimental environment, with frequent changes and the constant reindexation of the collection.

Regardless of the approach followed, we must have resources, techniques and tools to help with the translation, bearing in mind that the objectives of a CLIR translation are not exatly the same as an automatic traditional translation: in a CLIR system it is important that the translation does not lose the meaning of the original words, even to the detriment of the syntactic structure of the sentence and the introduction of noise. There is therefore a strong tendency not to rely on a single resource, but to integrate all available resources, such as MTs and MRDs, multilingual databases, parallel corpora and comparable corpora.

5. Evaluation

5.1 Senseval

The semantic ambiguity of the natural language is a difficult problem to solve. There is ambiguity in the spech, in syntax and in polysemic words. Knowing the exact sense of words helps in tasks such as the Information Retrieval, Machine Translation, Categorisation of Text etc.

There are currently many programs that automatically determine the sense of some words (Word Sense Disambiguation or WSD) in a determined context [22]. Many of them require the intervention of an expert, who writes the theoretical

rules of disambiguation of polysemic words [381]. Others use linguistic resources, such as MRDs and databases to help WSD programs. The main problem is to determine which is the best and SENSEVAL tries solves this problem.

SENSEVAL is a recent creation and there have been only two editions. The first SENSEVAL was in the summer of 1998. English, French and Italian were the languages chosen for the competition. Finally, in September, a workshop was celebrated in England. The second was in the spring of 2001, with the workshop in June and with a total of 12 languages in the competition. The competition has the following format. The organisation provides a corpus to be used by the participants for testing and training purposes. Then, an evaluation corpus is provided, with strict rules about the duration of the experiments before the results are submitted, so that in less than a month, each participant must offer their results.

SENSEVAL is also a generator of linguistic resources for disambiguation, as all results are made freely available when the competition is over.

5.2 TREC Monolingual System

The Text Retrieval Conference[8] (TREC) is sponsored by the National Institute of Standards and Technology (NIST) and the Information Technology Office of the Defense Advanced Research Projects Agency (DARPA) as part of TIPSTER program.

The purpose of TREC is to give support ot research in retrieval systems as provider of the infrastructure required for large scale evaluation of the methodology of text retrieval. TREC is also a forum where university and the industry meet to exchange ideas.

In each TREC, NIST prepares a collection of documents and queries, so that participants can use their IR systems on them, returning a sorted list of the recovered documents. The results of each participant are compared with the judgement of relevancy (the "right" answers) and the final classification is established. TREC finishes with a workshop where the participants compare their ideas. This forum is held annually and the results are published in the workshop's proceedings. TREC assembles the researches and developers of different systems and compares the results that are obtained in different tests, previously standardised and agreed by all. For researchers, TREC is one of the most prestigious conferences of IR and is constantly growing, not only in the number of participants, but also in the number of different tasks that are confronted. Since TREC-1 the effectiveness of IR systems has doubled.

5.3 CLIR Forums

5.3.1 *Cross-Language Evaluation Forum*

The Cross Language Evaluation Forum[9] (CLEF) is an activity of annual character and of European ambit, born in the year 2000 and co-ordinated by the DELOS

[8] http://trec.nist.gov/

[9] http://www.iei.pi.cnr.it/DELOS/index.html

Network of Excellence for Digital Libraries conferences, in collaboration with the NIST and the TREC Conferences. In fact, affinity between the CLEF and TREC is high, because objective of CLEF is, as with TREC, to establish a competition between groups in the area of IR, favouring, in this way, the fast assimilation of the progress made by participating groups in each edition. The key difference between the two competitions is the high speciality of CLEF in tasks concerned with the multilingual IR.

Although the natural ambit of the CLEF is the European Community, giving special attention to its official languages, the presence of teams from of America and Asia becoming more important. Languages such as Chinese and Japanese are also welcomed, although they are currently not included in the competition.

5.3.2 *NII-NACSIS Test Collections for Information Retrieval and Text Processing*

Is known that the IR is a task whose difficulty varies according to the language considered. Therefore, many of the techniques that are valid for certain languages are not valid for others. For example, the extraction of words that form a document is a more or less trivial task, if we think in English or romance languages, because the words are often separated with spaces. This is not the case in languages such as Chinese, whose words are not clearly defined.

NII-NACSIS Test Collections for Information Retrieval and Text Processing[10] (NTCIR) is a series of biannual workshops which have the aim of establishing a framework to promote progress in research into IR and equivalent tasks, such as the generation of summaries, information extraction, and multilingual IR. The NTCIR conferences have an Asian ambit. Specific oriental linguistic resources are given to the participating teams, in languages such as Chinese and Japanese. The first NTCIR workshop, held in 1998, had among other tasks a task with the aim of retrieving documents in English from a set of queries in Japanese. In the next edition, a task was added to obtain documents in Chinese from English queries.

References

1. Baeza-Yates, R. and Ribeiro-Neto, B. Modern Information Retrieval, ACM Press Books, New York (1999).
2. Ballesteros, L. and Croft, W.B. Resolving Ambiguity for Cros-language Retrieval. Proceedings of SIGIR'98 (1998).
3. Braschler, M. and Schäuble, P.A language-independent approach to european text retrieval. Carol Peters, editor, Proceedings of the CLEF 2000 Cross-Language Text Retrieval System Evaluation Campaign (2000).
4. Braschler, M., Ripplinger, B. and Schäuble, P. Experiments with the Eurospider Retrieval System for CLEF 2001. Carol Peters, editor, Proceedings of the CLEF 2001 Cross-Language Text Retrieval System Evaluation Campaign (2001).
5. Callan, J.P., Lu, Z. and Croft, W.B. Searching distributed collections with inference

[10] http://research.nii.ac.jp.ntcir

networks, In Proceedings of the 18th International Conference of the ACM-SIGIR'95 pp. 21–28. New York: The ACM Press (1995).
6. Chen, A., 2001. Multilingual Information Retrieval using English and Chinese Queries. Carol Peters, editor, Proceedings of the CLEF 2001 Cross-Language Text Retrieval System Evaluation Campaign, (2001).
7. Croft, B. What Do People Want from Information Retrieval? (The Top 10 Research Issues for Companies that Use and Sell IR Systems). Digital Libraries Research and Development Magazine, http//www.dlib.org (1995).
8. Davies, M. The Polyglot Bible. http://mdavies.for.ilstu.edu/polyglot/(1999)
9. Edwards, J. and Lampert, M. Talking Data: Transcription and Coding in Discourse Research. Lawrence Erlbaum Associates Publishers (1992)
10. Fellbaum, C. WordNet: An Electronic Lexical Database. E. by C. Fellbaum, The MIT Press (1998)
11. Fox, C. Lexical Analysis and Stoplists. In Information Retrieval: Data Structures and Algorithms,'Chapter 7, Prentice-Hall (1992)
12. Frakes, W. Stemming Algorithms. In Information Retrieval: Data Structures and Algorithms, Chapter 8, Prentice-Hall (1992)
13. Francis, W. Problems of Assembling and Computerizing Large Corpora. E. by S. Johansson (1982).
14. Francis, W. and Kucera, H. Frequency Analisys of English Usage. Houghton Mifflin (1982).
15. Franz, M., McCarley, J.S. and Roukos, S. Ad hoc and multilingual information retrieval at IBM. Ellen Voorhees and Donna Harman, editors, The Seventh Text Retrieval Conference (TREC-7). National Institute for Standards and Technology. Special Publication 500-242 (1999).
16. Grefenstette, G. Cross-Language Information Retrieval. E. by G. Grefenstette Kluwer Academic Publishers (1998).
17. Harman, D. Relevance Feedback Revisited. Proceedings of the 15th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval (1992).
18. Harman, D. Relevance Feedback and other Query Modification Techniques (1992).
19. Hausser, R. Foundations of Computational Linguistics., Springer-Verlag, Berlin Heidelberg New York, (1999).
20. Hull, D. and Grefenstette, G. Querying Across Languages: A Dictionary- Bases Approach to Multilingual Information Retrieval. the 19th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval (1996).
21. Hull, D.A. and Grefenstette, G. Experiments in multilingual information retrieval. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (1996b).
22. Kilgarriff, A. SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs. Proc. LREC 58-588 Granada 1998.
23. Krovetz, R. Viewing Morphology as an Inference Process. Proceedings of SIGIR'93 ACM Press (1993).
24. Lewis, D.D. Representation and Learning in Information Retrieval. Department of Computer and Information Science, University of Massachusetts (1992).
25. Lu, X.A., Holt, J.D., Miller, D.J., Boolean System Revisited: Its Performance and its Behavior, TREC-4 proceedings page 459, November (1995).
26. Luhn, H.P. The automatic creation of literature abstracts, IBM Journal of Research and Development. 2 (2) (1958).
27. Martinez, F., Ureña, L.A. and Garcıa, M. WWW como Fuente de Recursos

Lingüísticos para su Uso en PLN. Procesamiento del Lenguaje Natural (27) (2001).
28. Miller, G., Leacock, C., Randee, T. and Bunker, R.A. Semantic Concordance. Proceedings of the 3rd DARPA Workshop on Human Language Technology (1993).
29. Miller, G. WordNet: A Lexical Database for English. Communications of the ACM 38 (11) (1995).
30. Nie, J., Simard, M., Isabelle, P. and Durand, R. Cross-language information retrieval based on parallel texts and automatic mining parallel texts from the Web. In ACM SIGIR'99, pp. 74–81 (1999).
31. Oard, D. Cross-Language Text Retrieval Research in the USA. Presented at 3rd ERCIM DELOS Workshop, Zurich, Switzerland (1997).
32. Porter, M.F. An algorithm for suffix stripping. Program-automated library and information systems 14(3) (1980).
33. Rijsbergen, C.J. Information Retrieval. Butterworths (1979).
34. Salton, G. and McGill, M.J. Introduction to Modern Information Retrieval. McGraw-Hill (1983).
35. Salton, G. Automatic Text Processing: the transformation, analysis and retrieval of information by computer Addison Wesley (1989).
36. Savoy, J. Report on CLEF-2001 Experiments. Experiments with the Eurospider Retrieval System for CLEF 2001. Carol Peters, editor, Proceedings of the CLEF 2001 Cross-Language Text Retrieval System Evaluation Campaign (2001).
37. Sheridan, P., Braschler, M. and Schäuble, P. Cross-language information retrieval in a multilingual legal domain. Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries, pp. 253–268 (1997).
38. Small, S.L. Word Expert Parsing: A Theory of Distributed Word-Based Natural-language Understanding. Ph.D. Thesis, Department of Computer Science, University of Maryland, Maryland (1981).
39. Ureña, L.A. Resolución de la Ambigüedad Léxica en Tareas de Clasificaión Automática de Documentos, Ph.D. Thesis, Department of Lenguajes y Sistemas Informáticos University of Granada, Granada (2000).
40. Verdejo, F. Comprensión del Lenguaje Natural: avances, aplicaciones y tendencias en Procesamiento del Lenguaje Natural: fundamentos y aplicaciones. Documentacion del curso de verano de 1994 de la UNED, Ávila (1994).
41. Voorhees, E.M. Gupta, N.K. and Jhonson-Laird, B. The collection fusion problem. Proceedings of TREC'3, pp. 95–104. Gaithersburg: NIST Publication 500-225 (1995).
42. Vossen, P. EuroWordNet: A multilingual database with lexical semantic networks. Dordrecht: Kluwer (1998).
43. Yokoi, T. The electronic dictionary. Communications of the ACM 38 (11) (1995).
44. Zipf, G.K. Human Behavior and the Principle of Least Effort. Addison-Wesley (1949).