Experiences in Evaluating Multilingual and Text-Image Information Retrieval

Ana M. García-Serrano,[1,*] José L. Martínez-Fernández,[2,†]
Paloma Martínez[2,‡]
[1]*Artificial Intelligence Department, Technical University of Madrid,*
Campus de Montegancedo S/N, 28660 Madrid, Spain
[2]*Advanced Databases Group, Universidad Carlos III de Madrid,*
Avda. Universidad, 30, 28911 Leganés, Madrid, Spain

One important step during the development of information retrieval (IR) processes is the evaluation of the output regarding the information needs of the user. The "high quality" of the output is related to the integration of different methods to be applied in the IR process and the information included in the retrieved documents, but how can "quality" be measured? Although some of these methods can be tested in a stand-alone way, it is not always clear what will happen when several methods are integrated. For this reason, much effort has been put into establishing a good combination of several methods or to correctly tuning some of the algorithms involved. The current approach is to measure the precision and recall figures yielded when different combinations of methods are included in an IR process. In this article, a short description of the current techniques and methods included in an IR system is given, paying special attention to the multilingual aspect of the problem. Also a discussion of their influence on the final performance of the IR process is presented by explaining previous experiences in the evaluation process followed in two projects (MIRACLE and OmniPaper) related to multilingual information retrieval.

1. INTRODUCTION

As a result of the impressive evolution of the Internet, much effort has been put into developing information retrieval (IR) processes by trying to improve the user access to all information available online. Nowadays, the retrieval process is related not only to documents, that is, textual content, but also to multimedia information (images, audio, and video). Along with information and metadata formats, the multilingual dimension has become an important aspect to be taken into account.

*Author to whom all correspondence should be addressed: e-mail: agarcia@dia.fi.upm.es.
†e-mail: joseluis.martinez@uc3m.es.
‡e-mail: paloma.martinez@uc3m.es.

The user must be able to interact with the system using several languages, and, of course, documents in all these languages should be retrieved.

An IR process is all about the selection of documents from a collection to satisfy the information needs as stated by a user. The “high quality” of the output in an IR process is related to how good the selection is and to the information included in the retrieved documents. But, how can “quality” be measured? Notice that the quality of an IR depends on the user who is interacting with the system. Obviously, from a scientific point of view, this subjectivity must be suppressed to allow a comparison between different implementations of an IR process.

One important step during the development of an IR process is the evaluation of the output regarding the integration of different methods devoted to enhancing some specific aspects (tokenization, indexing, matching, etc.). Although some of these methods can be tested in a stand-alone way, it is not always clear what will happen when several methods are integrated. For this reason, almost all research tries to establish a good combination of several methods or correctly tune some of the algorithms involved. The current approach is not to evaluate the techniques applied separately, but to measure the precision and recall figures obtained when different combinations of techniques are included in an IR process.

Sections 2 and 3 include a short description of current techniques, paying special attention to the multilingual aspect of the problem. Also a discussion of their influence on the final performance of the IR process is presented. Then, previous experiences in the evaluation process followed in two projects (MIRACLE and OmniPaper) related to multilingual information retrieval are explained. Finally, some conclusions, including further research, are given.

2. EVALUATION IN INFORMATION RETRIEVAL PROCESSES

Traditionally, IR processes have been evaluated with the so-called *precision* and *recall* measures. *Recall* is the ratio between all relevant documents retrieved for a given query and all relevant documents existing in the collection for that query. On the other hand, *precision* is the ratio between the number of relevant documents retrieved and the total number of documents returned (relevant or not) for a given query. Both measures are related in such a way that the greater the precision the lower the recall and vice versa.

There are several well-known theoretical models that can be used to retrieve relevant documents by matching the query and the available documents[1]: The *Boolean Model*, the simplest one, is used by several of the search engines available on the Internet. It is based on the representation of documents by chains or sequences of bits. The *Vector Space Model*,[1] introduced by Salton, is based on an algebraic view of the IR problem. Documents and user queries are modeled using vectors that are depicted in an *n*-dimensional space, where each dimension is a word. The *Probabilistic Model*[2] was introduced by Robertson and Sparck-Jones, in which the final goal is to find the set of documents that maximize the probability for them to be relevant for a given user query.

To evaluate performance of processes applying these models, when separate precision and recall measures are considered, it is sometimes difficult to determine which one is better. To combine both figures, several measures have been defined.[1] One of them is the E-measure, which can be expressed using the following formula:

$$E = 1 - \frac{(1 + b^2)PR}{b^2P + R}$$

where P is the precision value for the process, R is the recall value, and b is a parameter used to promote recall against precision or vice versa. For example, if $b = 0.5$, recall is twice as important as precision. This measure tries to let the user decide whether precision is more important than recall or vice versa.

Another measure is the F-measure, which follows the formula

$$F_\alpha = \frac{1}{\dfrac{\alpha}{P} + \dfrac{(1 - \alpha)}{R}}$$

where the value of α is also used to make precision more relevant than recall or vice versa. Some other measures have been defined in an effort to cope with the IR process's ability to return relevant documents at the beginning of the ranked list. These IR evaluation measures have been widely criticized[3]:

(1) Regarding the size of the document collections used in the evaluation task, the main point is that these evaluation parameters have been obtained empirically, so even if good results are obtained for small and domain-specific document collections, it is not possible to generalize these results to large and open domain collections. On the other hand, obtaining precise recall and precision reference values for large collections is impossible because of the difficulty of the human judges to read a vast number of documents to make the relevance judgment (i.e., which documents must be returned for each defined query).

(2) Some of these works also argue that other factors have to be taken into account, such as the usability of the system, the degree of user satisfaction, and so forth.

To overcome these problems, there are several forums devoted to the definition of test collections, including documents, queries, and relevance judgments. These evaluation frameworks allow research groups to test their systems in a rigorous way without dedicating too much effort to the evaluation tasks and focusing only on the development and improvements in IR techniques.

TREC[a] is one of these; sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA), this conference has been held since 1992. TREC is in charge of building the collection of documents to be used (over 1 million documents), the set of

[a] Text REtrieval Conference, http://trec.nist.gov.

queries to be used, and the set of relevance assessments needed. With these data, it is possible to compare the effectiveness of different IR processes as well as Information Extraction (IE)[b] processes.

CLEF[c] was created[4] from the multilingual tracks defined in TREC. CLEF, sponsored by the IST Program of the European Union, is centered in the multilingual dimension of the IR process. This forum defines several *tracks*, each one with different goals. The data and the languages are changed in each edition. For the 2004 edition, the following tracks have been defined:

- Cross-Language Information Retrieval. Its goal is to retrieve documents in different languages using queries also written in different languages.
- Cross-Language Information Retrieval on Structured Documents (GIRT). This track is similar to the previous one, except that structured documents from a specific domain (social science data) are considered.
- Interactive Cross-Language Information Retrieval, iCLEF, devoted to including the user perspective in the evaluation process.
- Cross-Language Retrieval on Image Collections, ImageCLEF, devoted to retrieving images as answers to a user query using captions and the content of images.
- Cross-Language Spoken Document Retrieval (CL-SDR) especially designed to work with noisy transcriptions of audio recordings. Obviously, this kind of input introduces new challenges in the treatment of textual content (such as guessing speech boundaries and processing text with noise).

3. A SURVEY ON RESOURCES AND TECHNIQUES TO ENHANCE IR PROCESSES

From the simplest point of view, an IR process can be divided into three basic tasks: formulation, indexing, and comparison tasks. The formulation task is in charge of capturing and building a representation of the user information needs, the indexing task is related to obtaining a characterization of the documents, and the comparison task is devoted to the matching procedure between user queries and documents, returning a ranked list with those documents most likely to satisfy the user information needs (the so-called *relevant* documents to the query).

All these tasks have the retrieval model in which the system is based as a common factor, that is, to be able to implement the comparison task. Both query

[b] Techniques applied in IR are also used in other research areas such as IE. There is an important difference between IR and IE; whereas IR processes are devoted to the retrieval of complete documents, the main goal of IE processes is to obtain precise information from a document collection. This means to retrieve only words, small phrases, or parts of sentences that fulfill the information needs supplied by the user. For this purpose the IE process applies a number of statistical methods taken from IR, but the necessary enhancement comes from a deeper comprehension of the document contents, making use of different natural language processing (NLP) techniques. This content understanding has been the objective of the Message Understanding Conferences (MUC; http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html) held throughout the 1990s. These conferences defined a common framework to evaluate these kinds of systems and, nowadays, are part of the TREC conferences.

[c] Cross-Language Evaluation Forum, http://clef.isti.cnr.it.

and document representations must be built using the same theoretical model. To improve IR process efficiency, three approaches can be followed:

(1) Modify the matching task; that is, select another IR model or change the expression used to obtain similarity between query and documents.
(2) Act on the document characterization task; that is, based on a fixed model, the set of terms and weights used to characterize a document can be modified.
(3) Act on the formulation task; that is, choose different words or concepts to represent the information needs stated by the user in the form of a written query.

Moreover, concerning the types of resources used there are three main trends in the characterization of documents and queries and how it affects the information retrieval process:

- semantic approaches that try to implement some degree of syntactic and semantic analysis of queries and documents; this involves reproducing the understanding of the natural language text in a certain way;
- statistical approaches that retrieve and rank documents according to the match of documents–query in terms of some statistical measure; and
- mixed approaches that combine both of them trying to complement the statistical approach with semantic approaches by integrating NLP techniques and resources to enhance the representation of queries and documents and, consequently, to produce suitable levels of recall and precision measurements.

Throughout this section some of the most important techniques applied in the IR process are described, to show their importance in the final performance. Although some of these methods can be tested in a stand-alone way, it is not always known what is going to happen when several methods are applied sequentially. For this reason, almost every research work submitted to TREC or CLEF forums tries to establish a right combination or a correct tuning of some of the involved algorithms.

3.1. Indexing Task Techniques

According to points 2 and 3 stated in the previous paragraph, there are some techniques that can be applied at the indexing stage and other ones for the formulation stage. This subsection covers techniques to be applied at the indexing stage.

Tokenization: The first step in the analysis of a document is to split its content into individual words. This process is called *tokenization*. The output of the tokenization stage is used to measure statistics on the frequency of appearance of words in a document and, of course, in a collection of documents. The way in which words or combinations of them are recognized is a key issue in IR. It is important to know whether compound words, like the Spanish word “pelirroja” (*redhead*) are going to be divided into its constituents parts. Obviously, for this purpose specific information about the language in which the document is written is needed. For example, common words are joined into a single word in German, so a special algorithm to decompose words must be available. If languages in

Cyrillic, like Russian, or languages in other alphabets, like Japanese, are going to be tokenized, the problem becomes even more difficult to solve.

There are no predefined methods to evaluate the quality of the tokenization method applied in an IR process. It only can be guessed from the recall and precision values obtained for the whole system.

Stemming: Once the text is split into words, a process to obtain the stem of each word can be useful. The final goal in IR tasks is to recognize different concepts present in a document. Taking this into account, the stemming technique tries to build groups of words, each having the same stem. Each group will be represented by this common stem. The idea behind it is that words differing only in their terminal parts refer to the same concept, and only these concepts are of interest in an IR task. Stemming algorithms do not have a 100% precision; in some situations wrong stems are obtained.[d] As can be seen, stemming has an important influence on the computed distribution of words in a document. There are algorithms, available in different languages, that are able to obtain word stems applying only morphological information. The best known is the Porter algorithm,[5] developed for the English language. Some different versions of this algorithm for other languages are available as part of the Snowball tools.[6] Several research works[7] have been devoted to proving that stemming leads to better precision–recall figures in IR systems, which is a valid assumption in monolingual environments.[8]

Lemmatization: In a way similar to stemming, this technique is devoted to the clustering of words according to their lemma, that is, taking into account a canonical representative for all words differing only in their terminal parts. For example, the Spanish verbs “corro,” “corres,” “corren,” “corremos” (*I run*, *he runs*, *they run*, *we run*) will be assigned to the same lemma, “correr” (*to run*), the infinitive form for the verb. The canonical representative of a word cannot be obtained by applying some kind of algorithm, as in stemming. The lemmatization task requires linguistic resources (usually lexical databases) for the target language, an important issue to take into account in multilingual environments. Of course, because not every word is covered by these lexical databases the lemmatizing task cannot achieve a 100% precision. There are no well-established methods to measure the quality of these resources. Usually the number of words that is managed in the lexical resource is the reference figure, but the domain covered by the resource must also be taken into account.

Shallow parsing: An alternative to the tokenization process is to carry out a shallow linguistic analysis of the input text. This linguistic analysis includes morphologic and syntactic information on the input text. In shallow parsing, not only are tokens recognized, but also information on gender, number, and class (noun, verb, adjective, etc.) for each word is obtained and then used to identify groups of words in sentences according to their linguistic function. Knowledge of the morphosyntactic structure of a document can lead to more precise information on the semantic content of the document. This kind of analysis is computationally expensive, so it is not usually applied in IR processes, but is needed for IE and

[d] This approach leads to problems of “understemming” and “overstemming.”

question-answering tasks. The evaluation of these techniques involves manually labeled document collections, where morphosyntactic information is present and supplied by human experts.

Entity recognition: Proper nouns or combinations of words referring to concrete entities can be very useful in an IR task. Words used to represent these entities must be grouped and can be treated as a unit, because they are supposed to have greater discrimination capacity than simple words appearing in a document. Results of the shallow parsing task can be used to identify entities present in a document. If no morphosyntactic information is available, some algorithms based on simple heuristics can be used to recognize proper nouns in a text. To be able to detect classes of entities (person, organization, etc.), some external resources would also be needed. Again, the multilingual dimension makes it necessary to have specific lexicons for each language, which are very difficult to maintain. The evaluation of these algorithms and techniques is a complex task, since there are no specific resources and discussion forums devoted to them. When MUC conferences were held, entity recognition systems could be tested but, nowadays, the interest in these techniques has either waned or the problem of recognizing these structures is solved.

Statistical methods: IR processes rely on the following idea: The greater the number of times a word appears in a document, the greater is its importance in characterizing that document. Some nuances can be added to this idea: The length of the document is important, so frequencies of appearance for the words are usually normalized according to the total number of words present in a document. On the other hand, given a collection of documents, the ability for a word to identify a subset of them depends on the number of documents in the collection in which this word appears. For example, if a word appears in all the documents of a collection, it is not useful to identify a subset of documents. These figures correspond to the so-called term frequency for a term in a document (tf) and the inverse document frequency for a term (IDF). It can be seen that resources are needed to make this kind of analysis non-language-specific. This is the great advantage of statistical approaches to the IR problem: It is the same for every language (assuming that it is possible to make a correct token splitting of the text). Another statistical approach that is becoming of greater interest is the one based on *n-gram* detection. A *gram* is a chain of characters with length *n*. This chain is built by taking a window of *n* positions, which is shifted over the input string. The final result is a set of smaller character chains of size *n* that are used to represent the input string. In this way, no complete words are indexed, only their *n-gram* form. If this approach is carried out, no linguistic techniques can be applied and, obviously, no specific resources are needed. Besides, there is no need to include adaptations to take into account multilingual environments, except if Asiatic languages are considered. There are works[9] where *n-grams* are not only used in the IR process, but also for machine translation (MT) processes if parallel corpora are available.

3.2. Formulation Task Techniques

According to the description given for an IR process, documents and user queries must be represented according to the same common model. This implies that the same analysis methods applied to documents must also be used to represent the user query. If documents have been indexed from the stems of the words appearing in the text, the user query must also be expressed using the stems of the words supplied. But there are still some methods aimed at improving the representation of the query. Descriptions of these methods are given below.

Query expansion: Sometimes it can be a difficult task for the user to provide the same word in the query as the one appearing in the document collection. The main goal of an IR process is to deal with concepts instead of concrete words but, frequently, the same concept can be denoted using different words (synonyms). For example, the Spanish words "empresa" (*firm*) and "compañía" (*company*) are synonyms. Let us suppose that in the document collection only the first word appears and the user supplies the second one in the query. In this situation no documents could be returned for the user query "compañía," but there are documents about "empresa," which can relevant for the user. To avoid these undesired situations, the user query could be expanded with related terms, adding synonyms of the query words. There is no automatic way of obtaining the synonyms of a given word, so specific lexical resources must be made available. WordNet[10] is an ontology developed for the English language where words with the same meaning are grouped into structures called *synsets*. In addition, other relationships between these synsets are stored, such as hyponym and homonym relationships. There is a multilingual version of this ontology, EuroWordNet[11] where the lexical database is broadened to cover other languages, including equivalence relationships between synsets in different languages.

Related to the use of lexical resources, ambiguity is a crucial problem to be solved in IR processes. Although lexical resources are available, the issue of adding related terms to user queries is not a trivial task; it is necessary to consider the context of words in queries in order to remove unsuitable synonyms. Research work presented in Ref. 12 shows that dealing with lexical variation is more beneficial for incomplete and relatively short queries; under the assumption that every synonym is added to a query word and if no disambiguation task is carried out, it was expected that the retrieval process itself would carry out a disambiguation process because a conjunction of terms would eliminate many of the spurious forms.

Nevertheless, there are some doubts about the effectiveness of this technique to improve precision and recall in IR processes. Although some research works[12] obtained better results, other research works[6] could not benefit from query expansion. This fact can be a result of the application domain covered by these projects.

Relevance feedback: Continuing with the ambiguity problem, in some situations a single word can be used to represent different concepts (polysemy). An analysis of the queries written by users in web search engines can reveal that a query string comprises only a few words,[13] and there is not enough information to carry out a disambiguation task. For example, in the Spanish word "banco" (*bank*) can refer to a bank (financial entity) or a bench (a seat in the park); so, if the user supplies this word, which documents should the IR process retrieve, those related to financial entities or those about seats in parks?

This disambiguation process can only be solved with the help of the user. So, an interactive process is defined to improve ranked lists provided by the system.

The procedure followed begins with a query expressed by the user. A ranked list of documents is returned and the user marks which of them are relevant. The system analyses these relevant documents and builds a new query by adding the most relevant terms appearing in these documents. The new query is then used to carry out a new search within the collection. This iterative procedure continues until the user is satisfied with the results obtained.

In some situations it is not possible to ask the user and the system must reproduce the previous behavior by itself, taking the first retrieved documents as relevant and using them to build a new query. This procedure is called *blind relevance feedback*.

3.3. Specific Tools to Manage the Multilingual Dimension

Throughout the former description of the different techniques applied in IR processes, the multilingual dimension has been considered, but machine translators are usually needed. The IR process within multilingual document collections can be based on three different approaches: First, the user query is translated into all languages appearing in the target collection; each translated version of the query is executed against the subset of documents in the same language and independent result lists are merged in some way; second, all documents are translated to the language used in the query; third, a single query is compiled with the translations of the initial query expression; this multilingual query is then executed against the whole document collection. Each approach has drawbacks and advantages but the third one is the most applied because no merging of language-dependent result lists is needed, as in the first approach. Several algorithms have been applied to the problem of merging result lists obtained for each language: *normalization*, where a single list is compiled by normalizing (according to the number of documents in each collection) the relevance value assigned to a document in a partial result list; *round-robin*, where the position of documents in each partial result list is considered (e.g., if three partial result lists are considered, then the first element of each partial list is taken to make up the first three documents of the final list, then the second element of each partial list is used to obtain the third to sixth positions in the final list, and so on); finally, the *unique multilingual index* technique is based on the construction of a single index for all documents, without taking into account the language used to write them; the retrieval process is executed on this single index, thus obtaining a unique result list. In Ref. 14, these methods are explained and a new approach is proposed and tested. This new approach, called *2-step RSV*, is based on building a new collection made up of the first L $(1 \leq N \leq 1{,}000)$ documents appearing in the partial results lists and executing the query against this new collection.

When comparing monolingual systems with multilingual ones, a decrease in recall and precision is usually seen. This is because of errors introduced by the translation tools. Some words have more than one translation in other languages, and it is not easy to select the right one. On the other hand, commercial products are used when available, but, depending on the languages involved, it is not always possible to find products able to cope with them. In this situation, some techniques, such as working with parallel corpora, can be applied, but the results are not as good as those desired.[15]

4. EXPERIENCES IN THE EVALUATION OF INFORMATION RETRIEVAL PROCESSES

This section deals with several research projects where some of the previously described techniques have been developed and tested. For this testing purpose a framework for the evaluation of IR processes has been applied.

4.1. MIRACLE at CLEF

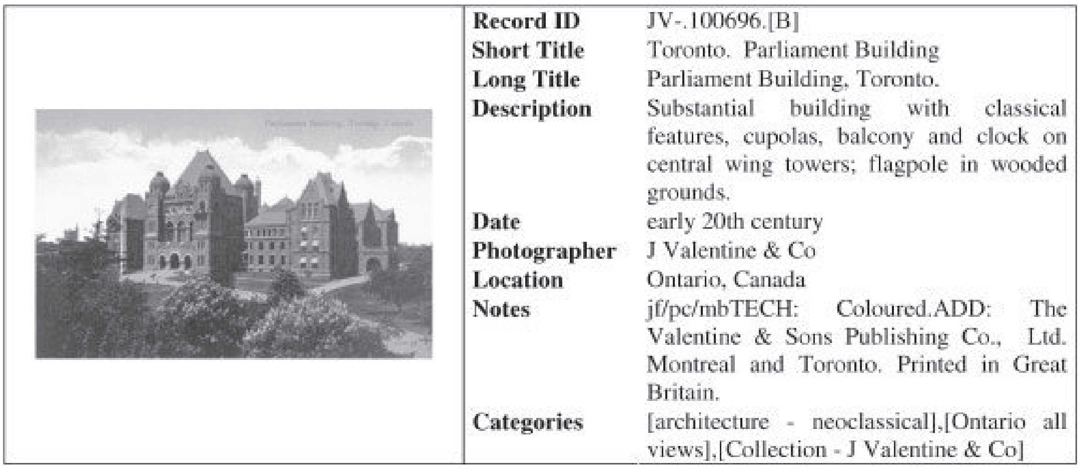
The Multilingual Information RetrievAl for the CLEf campaign (MIRACLE) research group is a team made up of professionals from several Spanish academic and industrial institutions. These are: Daedalus,[e] a leading company in linguistic technology in Spain, the Universidad Politécnica de Madrid (UPM), the Universidad Autónoma (UAM), and the Universidad Carlos III de Madrid (UC3M). This research group was created to share knowledge of linguistic processing and multilingual information retrieval and to work together in the construction of a system to take part in the CLEF campaign; the group has taken part in the 2003 and 2004 campaigns.

Tools provided by the organization include:

- Sets of documents, made up of newspaper articles covering different languages but within the same time period for each language.
- Sets of queries, written in different languages and having a predefined structure divided into three fields: title; a few basic words for the query; and narrative, a long description of the kind of content that an article must have in order to be returned by the system. Queries for cross-lingual tasks also include a description field, where a short sentence is used to point out the main subject of the query.
- Relevance judgments, sets of manually constructed files with information about the articles that should be returned by a system as part of the answer to each predefined query. Document collections supplied by the CLEF organization are of a considerable size, nearly 2 million articles taking into account all languages. So, these relevance judgments are obtained by a sampling process, where the intersection of the first 1,000 results submitted by each participant are considered as the set of relevant documents, and part of them are manually verified. A detailed description of the process can be found in Ref. 16.

The call for participation of CLEF is divided into several tracks, according to different possible environments where IR systems can be of any use. For a given track, several tasks can be defined according to different parameters, such as the languages that can be used or the kind of process, manual or automatic, used to build a query. Each research group taking part in CLEF can send several sets of results for each task, called *runs*. In some of the tasks, the number of runs is limited.

[e] http://www.daedalus.es.



Record ID	JV-.100696.[B]
Short Title	Toronto. Parliament Building
Long Title	Parliament Building, Toronto.
Description	Substantial building with classical features, cupolas, balcony and clock on central wing towers; flagpole in wooded grounds.
Date	early 20th century
Photographer	J Valentine & Co
Location	Ontario, Canada
Notes	jf/pc/mbTECH: Coloured.ADD: The Valentine & Sons Publishing Co., Ltd. Montreal and Toronto. Printed in Great Britain.
Categories	[architecture - neoclassical],[Ontario all views],[Collection - J Valentine & Co]

Figure 1. Image example from the St. Andrews historic collection.

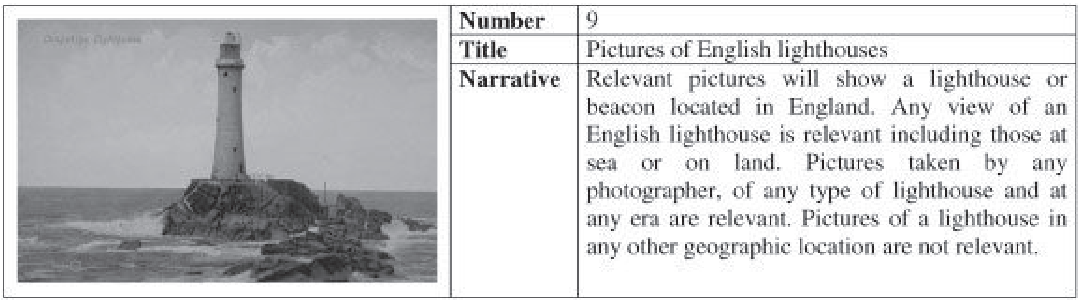
In this article, attention is centered on the MIRACLE participation in the ImageCLEF track where the main goal is to carry out multilingual searches within image collections using the content of the image and/or textual descriptions supplied for each one. In Figure 1 an image corresponding to the historical collection is shown. This is called the “St. Andrews image collection,”[f] a set of 28,133 images from the St. Andrews University Library image collection. Predefined queries, also called topics, have a specific structure, which is shown in Figure 2. Eleven languages were considered in query captions (Russian, Dutch, Swedish, Italian, Danish, German, Japanese, French, Chinese, and Finnish) and only the title field for the captions of the queries is translated into each language. For the purposes of the ImageCLEF track, manual and/or automatic runs can be submitted. The search process can be an iterative one where several search cycles are concatenated. The images used in each iteration can be manually selected or automatically determined, distinguishing between manual and automatic submissions. There is another medical image collection, which comes from donations from the University Hospitals of Geneva and is made up of scans and X-ray images along with short textual medical case descriptions in English and French.

4.1.1. Image Search System Architecture

The text-based image retrieval system is divided into several independent modules, in an effort to promote flexibility and adaptability in the architecture of the system. The modules are: a search engine, Xapian,[g] a freeware and general purpose search engine that is used to index and search the provided collection of image captions; a stemming module, in charge of extracting the stem of a word; the implementation used is based on the Porter Algorithm[5]; a stopword filter, used

[f] http://ir.shef.ac.uk/imageclef2004/stand.html.

[g] The Xapian project, http://www.xapian.org.



Number	9
Title	Pictures of English lighthouses
Narrative	Relevant pictures will show a lighthouse or beacon located in England. Any view of an English lighthouse is relevant including those at sea or on land. Pictures taken by any photographer, of any type of lighthouse and at any era are relevant. Pictures of a lighthouse in any other geographic location are not relevant.

Figure 2. Query example for the St. Andrews historic collection.

to remove empty words semantically, such as articles, prepositions, and so forth; the implementation uses a list of words to be ignored, which must be available for each language considered[6]; a tokenizer module, in charge of dividing the text into sequences of characters that can be considered as words in the target language; a proper noun detection module, developed by applying some basic morphosyntactic rules, used to detect the appearance of proper nouns in the texts provided; a probabilistic morphosyntactic tagger, adopted from Brill's work,[17] provides morphological and syntactical information on texts written in English; the output of this tagging component can be introduced as the input of a shallow parsing component, to detect linguistic groups in sentences that can lead to a finer characterization of the text. For semantic and translation purposes, EuroWordNet[11] has been integrated in the system, not only for synonym expansion of the query but also for query word translation purposes, by the Inter-Lingual Index provided by this resource. Finally, web translation tools[h] are applied when languages not covered by EuroWordNet are considered.

All described components can be applied in different combinations and configurations to define diverse sets of experiments, designed to improve the performance of the image retrieval process.

4.1.2. Design of Experiments

Each experiment carried out is given a unique name by trying to summarize its main characteristics. There are three main types of experiment: monolingual, where the query and the image captions are written in English; bilingual, where the query language is other than English; and text based plus content based, where partial lists obtained with content-based and text-based search systems are mixed. The structure for the monolingual English run names is shown in Figure 3. Figure 4 shows the names structure of bilingual experiments.

Table I summarizes the different experiments defined for the task according to the aforementioned three types of experiment. For the monolingual

[h] Translation Experts, http://www.transexp.com; Altavista's Babel Fish Translation Service, http://babelfish.altavista.com/.

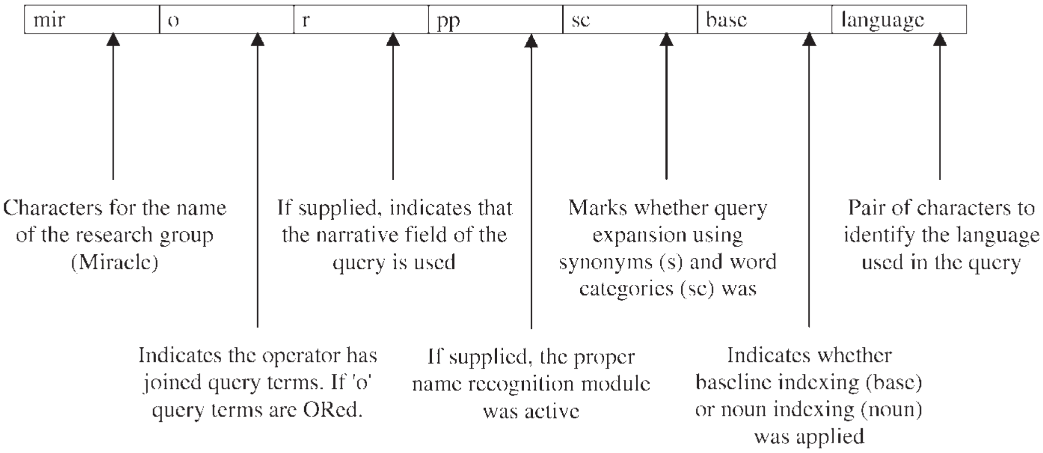

mir
o
r
pp
sc
base
language
Characters for the name of the research group (Miracle)
If supplied, indicates that the narrative field of the query is used
Marks whether query expansion using synonyms (s) and word categories (sc) was
Pair of characters to identify the language used in the query
Indicates the operator has joined query terms. If 'o' query terms are ORed.
If supplied, the proper name recognition module was active
Indicates whether baseline indexing (base) or noun indexing (noun) was applied

Figure 3. Monolingual run name structure.

experiments, specific linguistic techniques have been applied, which include the following:

- *WordNet*[10] *query expansion module*, in charge of obtaining the synonyms for a given word; these synonyms can be filtered out by their linguistic category (noun, verb, adjective, etc.), that is, only synonyms in the same linguistic category as the initial word are included. This linguistic category filtering is applied in an attempt to reduce the ambiguity introduced by the synonym expansion. Experiments using expansion through synonyms are marked with an "s" in the run name. If the linguistic category for the word is filtered out, an "sc" letter combination is included in the run name.
- *Proper names detection module*, used to automatically recognize proper names appearing in the image captions and in the queries. The idea behind this technique is based on the belief that proper names give more information to characterize a document than a regular word. If an experiment uses this component, the string "pp" is included in the run name.

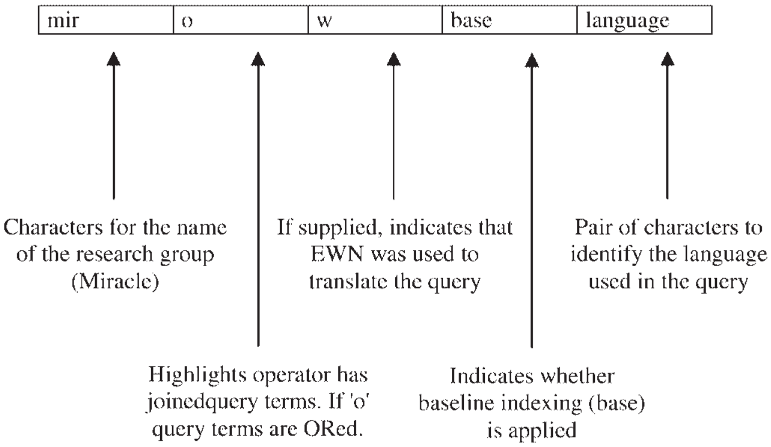

mir
o
w
base
language
Characters for the name of the research group (Miracle)
If supplied, indicates that EWN was used to translate the query
Pair of characters to identify the language used in the query
Highlights operator has joinedquery terms. If 'o' query terms are ORed.
Indicates whether baseline indexing (base) is applied

Figure 4. Bilingual run name structure.

Table I. Types of experiments defined for ImageCLEF 2004.

Kind of experiment	Description	Run names
Monolingual	English queries against the English collection. Use of WordNet[10] for query expansion, proper name detection, and lexical information to support query expansion through synonyms.	mirobaseen mirosbaseen mironounen mirosnounen miroscnounen miroppbaseen mirosppbaseen miroppnounen mirosppnounen miroscppnounen mirorppbaseen mirorscppbaseen
Bilingual	EuroWordNet is used for translation purposes. For languages not covered by EWN, online translators are used.	mirobaseru mirobasedu mirobasesw mirowbaseit mirobaseda mirowbasees mirowbaseesc mirowbasege mirobaseja mirowbasefr mirobasezh mirobasefi
Text-based + Content-based	Results for monolingual experiments are used in combination with GIFT 0.1.9,[18] a content-based image retrieval tool.	enenrunexp1 enenrunexp7 enenrunexp4 enenrunexp10

- *Lexical information module*, which can produce a morphosyntactic analysis of a text. This analysis returns the category in which a word has been used in the text. In this way, words can be selected to be part of the characterization of the document according to their linguistic category. An implementation of the Brill tagger[17] was applied for this purpose. In the experiments defined, if only words acting as nouns are used to index the document collection, the "noun" mark appears in the run name.

In the cross-language experiments, it was not possible to carry out the same experiments as those defined for the monolingual ones. This was as a result of the lack of resources to make linguistic analysis for every considered language other than English. As an alternative, two different approaches where applied, depending on the availability of linguistic resources for the involved languages. In the first one, EuroWordNet interlingual index (ILI) is used to translate the query terms. EuroWordNet was used for licensed languages such as German, French, Spanish, and Italian. In the second one, publicly available web translators, Systran[i] and Translation Experts[j] were used to translate the query terms.

The last set of experiments defined in Table I comprises runs where content-based image retrieval is mixed with text-based image retrieval. The purpose of these experiments is to take the best of both methods for image retrieval. The process followed to combine these two approaches begins with the execution of a text-based search, obtaining a list of results by searching through image captions. The first N (where N is a configuration parameter) results of this list are used to build a query for the content-based image retrieval (CBIR) system.

The CBIR system implementation[18] allows the definition of a relevance feedback task to improve the performance of the retrieval process. This relevance feedback task consists of a new search using the first M results of an initial search. For the experiments defined in this work, only the first five elements are used for this purpose (so $M = 5$). Partial result lists are then merged according to Expression 1:

$$\begin{cases} \sqrt[k]{REL\_VIS^{weight\_vis} \times REL\_TXT^{weight\_txt}}, & \begin{array}{l}\text{for elements in both lists and} \\ k = \text{weight\_vis} + \text{weight\_txt}\end{array} \\ \text{factor\_vis}, & \begin{array}{l}\text{for elements appearing only} \\ \text{in the list obtained with the} \\ \text{CBIR subsystem}\end{array} \\ \text{factor\_txt}, & \begin{array}{l}\text{for elements appearing only} \\ \text{in the list obtained with the} \\ \text{textual search subsystem}\end{array} \end{cases} \tag{1}$$

In this formula, *REL_VIS* is the relevance figure returned by the CBIR system, *REL_TXT* is the relevance value returned by the text-based image retrieval system, *weight_vis* is a parameter to highlight the importance of the content-based result, *weight_txt* is a parameter to highlight the importance of the text-based result, *factor_vis* is a constant relevance value given to elements appearing only in the content-based partial results list, and *factor_txt* is the counterpart of *factor_vis* for the text-based system. Figure 5 depicts the process followed in this type of experiment.

In Table I, experiments relative to the merging of content-based and text-based retrieval methods used some of the previously defined text-only runs, specifically *mirobaseen*, *mirosppbaseen*, *mirosnounen*, and *miroscppnounen*.

4.1.3. Experimental Results

Table II shows evaluation results obtained for the previously described experiments in the ImageCLEF 2004 campaign. The “Rank” column in the table gives

[i] http://babelfish.altavista.com.
[j] http://www.transexp.com.

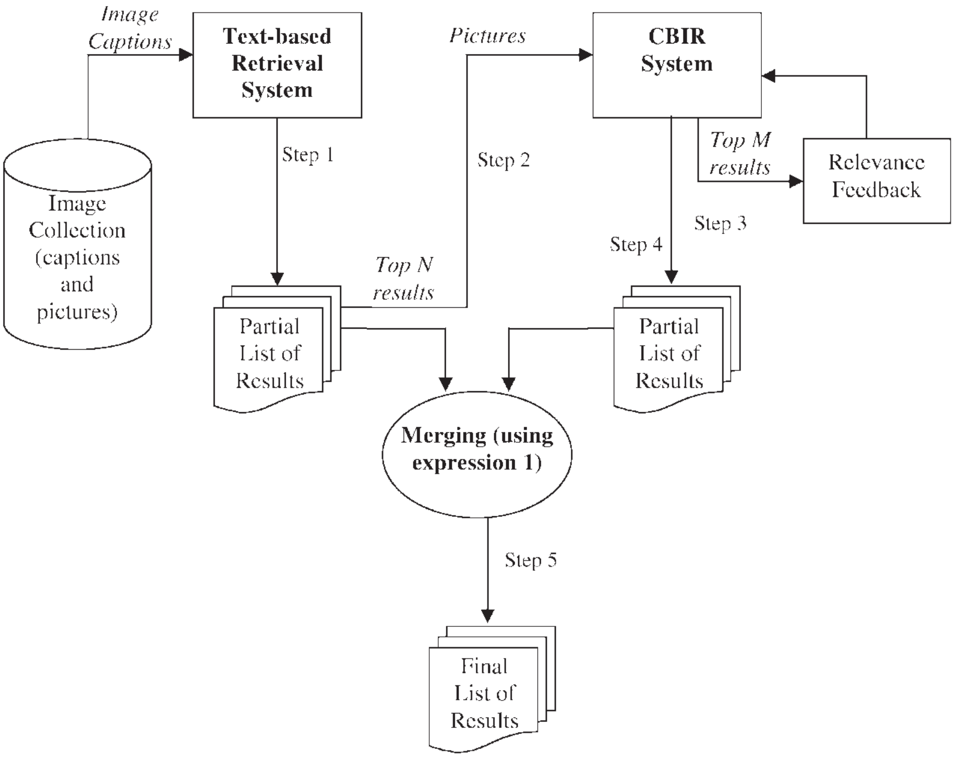


Figure 5. Architecture to combine text and content-based image retrieval.

the absolute place of the corresponding run in the list of all submitted runs ordered by decreasing mean average precision. This mean average precision is included in the "MAP" column. The "% Monolingual" column represents the distance between bilingual and monolingual experiments; for example, for the *mirobaseru* experiment, the MAP obtained is 0.3866, which is a 65.93% of the best monolingual MAP obtained (taking into account all kinds of experiment).

Results provided in Table II show a great gap between the monolingual and multilingual experiments, around a 34% decrease in average precision. In the ImageCLEF 2003 campaign, bilingual experiments were defined for French, German, Spanish, and Italian, where the average precision was 10% better than the average precision obtained in the 2004 edition. Taking these facts into account, it seems that EuroWordNet, when used for translation purposes, is not as good as online web translators. On the other hand, there were participants who obtained only a 10% decrease in precision between the best monolingual and the best bilingual experiments, but they were not using EuroWordNet as a translation tool. If attention is paid to the experiments where content-based and text-based techniques are merged, small variations between the text-based result and the final results can be seen. This means that the content-based part of the system does not improve (or decrease) retrieval performance, but more tests must be carried out to obtain a decisive conclusion.

Table II. Average precision figures for each submitted run.

Run name	MAP	% Monolingual	Rank
mirobaseen	0.5865	NA	1
enenrunexp1	0.5838	NA	2
mirosbaseen	0.5623	NA	4
miroppbaseen	0.5609	NA	6
mirosppbaseen	0.5388	NA	8
enenrunexp7	0.5339	NA	9
mirobaseru	0.3866	65,93	73
mirobasedu	0.3807	64,91	76
miroppnounen	0.3384	NA	87
mirosnounen	0.3383	NA	88
enenrunexp4	0.3373	NA	89
mirorppbaseen	0.3366	NA	90
mirosppnounen	0.3337	NA	92
mirobasesw	0.3043	51,89	99
mirowbaseit	0.2857	48,72	106
mirobaseda	0.2799	47,72	107
mirorscppbaseen	0.2703	NA	112
mirowbasees	0.2687	45,82	113
mirowbaseesc	0.2615	44,59	114
miroscppnounen	0.2568	NA	116
enenrunexp10	0.2533	NA	118
mironounen	0.2525	NA	119
miroscnounen	0.2461	NA	120
mirowbasege	0.2455	41,87	122
mirobaseja	0.2358	40,21	124
mirowbasefr	0.2188	37,31	127
mirobasezh	0.1777	30,30	135
mirobasefi	0.17	28,99	141

4.2. OmniPaper: The Smartest European News Finder

The main goal of the OmniPaper[k] project is to define an approach to a smart access to news from different newspapers and in different languages. The key objective of the project is the creation of a multilingual navigation and linking layer on top of distributed information resources. As a final result, OmniPaper will set an overall entrance to the news repositories and, thus, tackle the problem of overinformation in accessing news articles as well as the problems coming from an unassisted search carried out by the user in specific newspaper sites, each often with their own user interface and search method. This entrance gate will allow the user to query the different databases (i.e., the sets of articles available at the sites of different news providers) with one single search in one language, thus without having to know the languages of the different archives.

In the OmniPaper project the mixed approach described in Section 3 is adopted: First, statistical methods are considered and then semantic techniques

[k] IST-2001-32174, www.omnipaper.org.

complement the statistical framework through some kind of syntactic and semantic processing carried out on both the news and user queries, but in a shallow way (not attempting to understand the text). This approach requires the availability of linguistic resources for every language involved in the project; that is, semantic approaches are language and domain dependent. The resources considered are: stemmers for reducing words that differ only by suffixes from the same root; proper-name heuristics for recognizing entities (personal, geographical, and institutional names); semantic resources (such as EuroWordNet) for enriching query terms, and bigram multiword detection.

The OmniPaper architecture includes the Automatic Keyword Extraction (AKE) module, which is in charge of the characterization of documents and queries using the Vector Space model. For evaluation purposes, three parameters were identified, one for each linguistic technique (i.e., stemming application, proper name detection, and bigram subdivision), and used to define the experiments to be carried out. The document collection used in the evaluation process was a set of 56,000 articles, approximately, taken from the Glasgow Herald. This collection is part of a document collection provided by the CLEF organization, and the set of queries and relevance judgments used to evaluate the AKE prototype was also part of this collection.

In the experiments defined for the evaluation of the AKE module (see Table III), the query language and the target document collection language were always English. Nevertheless, the AKE module is able to work with all the languages considered in the OmniPaper project, which are Catalan, Spanish, German, French, Dutch, English, and Portuguese.

The queries used for evaluation present a structure divided into three fields: a title, with some words on the main subject of the query; a description, one or two sentences about the query subject; and a narrative, where several paragraphs are provided to give a detailed description of the kind of information that is considered as relevant for the query. Figure 6 details the results obtained when only the title of the query is used to instigate the search. Figure 7 shows the results when the title and the description fields for the query are applied, and Figure 8 shows the situation when all the three query fields are used.

Some conclusions can be highlighted by taking these graphs into account. First, the use of long queries improves precision and recall values. As can be seen,

Table III. Description of experiments defined for AKE evaluation with the CLEF data set.

Experiment name	Stemming application	Proper name detection	2-Word gram (bi-gram) subdivision
normal	No	No	No
stemming	Yes	No	No
nouns	No	Yes	No
ngrams2	No	No	Yes
stem_nouns	Yes	Yes	No

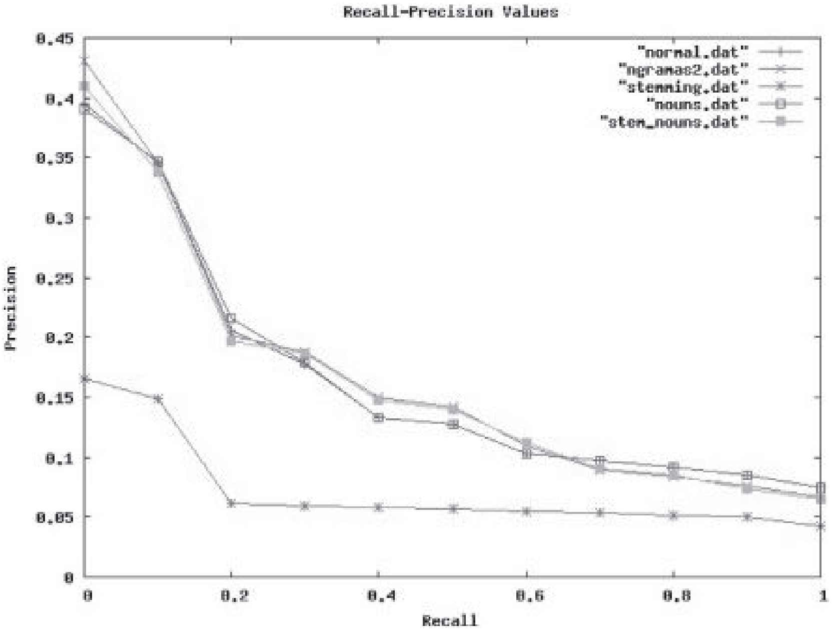


Figure 6. Evaluation results when only the Title field of the query is used.

when both title and description fields are used, precision acquires the maximum value. It is worth mentioning that increasing the query length by adding the narrative fields does not produce a substantial benefit. Second, the use of stemming improves results when medium-size queries are presented to the system. These results strengthen previous experiments where stemming has been applied.[19–23] A

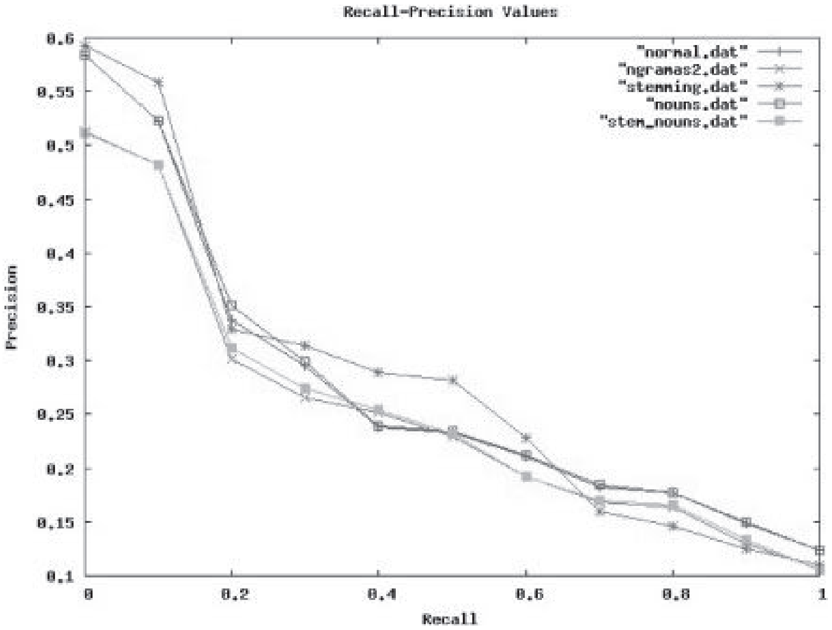


Figure 7. Evaluation results when the Title and Description fields of the query are used.

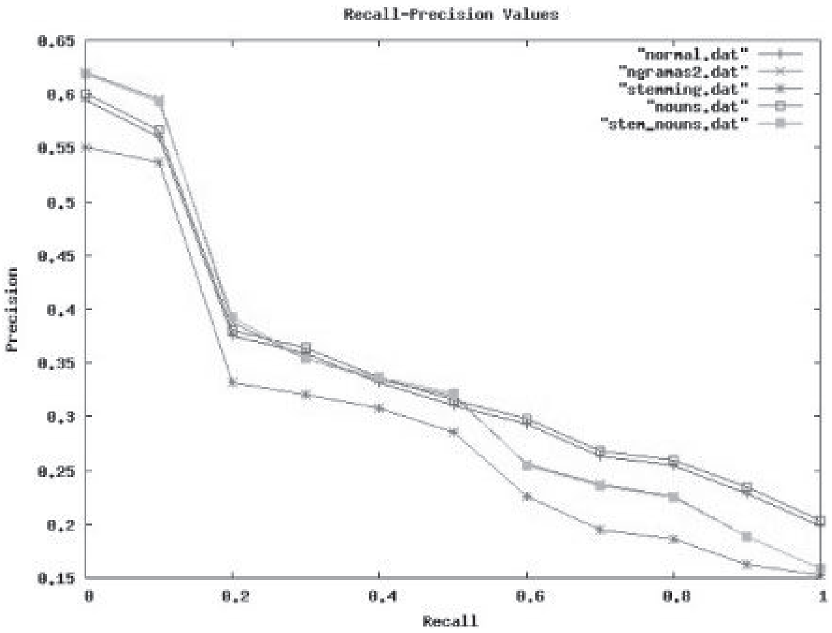


Figure 8. Evaluation results when the Title, Description, and Narrative fields of the query are used.

similar conclusion can be drawn from experiments where proper noun detection has been applied. In this situation, it is worth analyzing further the use of more refined techniques, such as entity recognition systems. Finally, if attention is paid to the use of statistical techniques such as *n*-grams, no improvement is found when bi-gram detection is activated. Further evaluation would be needed to discard the use of this kind of technique.

Before working with the formal test collection provided by CLEF, the Omni-Paper consortium developed a small test set. This set was made up of 1,881 English news articles, supplied by one of the news providers involved in the project. Along with these articles, 18 topics for different domains were developed by different members of the consortium. These topics where written along with the list of articles, manually determined, that should be retrieved for each topic. As can be seen, the way of building the test collection is very different than the procedure followed by the TREC/CLEF organization. These organizations use several human judges, whose relevance assessments are joined and compared. Larger document collections are also considered, and very carefully defined procedures are applied to determine the required sets of relevant and nonrelevant documents. The evaluation results obtained with the test collection developed by the consortium were not comparable to the results obtained with the CLEF data set and shown in Figures 6, 7, and 8.

5. CONCLUSIONS

The experiences described in the evaluation of IR systems show the huge amount of work required to supply good tools to allow the comparison and proper evaluation of these kinds of systems. The document collections provided must be big enough (between 200,000 and 500,000 documents per language) to ensure a possible real application of the systems. Forums devoted to this task define a common framework for researchers in the IR and IE fields to test their algorithms, methods, systems, and so forth. On the other hand, researchers must be careful with the scope of these evaluation frameworks. If developed processes only take tasks proposed by these forums into account, there is an obvious risk of narrowing their field of application. Besides, real applications are also concerned with the time needed to carry out the retrieval process, but, at the time of writing, this is not one of the evaluation parameters considered by the aforementioned forums. It is not necessary to say that the response time of an IR system is a crucial factor in a real application.

Throughout this article different techniques applied in IR processes have been described, focusing on the process evaluation and taking the presence of several languages into account. Two projects related to IR have been described, paying special attention to the evaluation procedure followed in each one. From our point of view, a general and interesting result must be highlighted: The application of sophisticated linguistic resources to the IR process does not always lead to better precision and recall figures. A standardized and well-defined way of applying this kind of information in a profitable way to the IR process has still not been discovered.

These experiences lead us to conclude that the existence of evaluation forums that provide material (documents, queries, and relevance assessments) needed to evaluate an IR process is crucial for the development of this field of research. Thus, it is also possible to evaluate NLP and other techniques adapted to IR goals.

Moreover, the availability of a suitable evaluation framework is crucial to allow working with different approaches to the IR problem and, especially, in order to evaluate the incidence of using resources, algorithms, and other things on documents in dealing with diverse domains and written in distinct styles. There have been some initiatives devoted to the development of platforms for helping in the development of applications needing NLP abilities, for example, the General Architecture for Text Engineering (GATE),[24] developed by the Sheffield NLP Group. This platform covers the whole life cycle of NLP components, including the evaluation step. This kind of platform is a good starting point for groups or companies taking their first steps in the NLP area, but, from the point of view of the authors, the evaluation phase of the process should be treated in isolation, developing standards and tools to support the evaluation of different NLP-based systems.

The research work presented in this article is being continued in several ways. First of all, the OmniPaper project has a continuation in the NEDINE[1] (Intelligent News Distribution Network for multinational business news exchange and dissemination) project. NEDINE's main goal is to provide a network to facilitate the integration of different European information sources, crossing the linguistic barriers. So, the NEDINE project is extending the techniques developed in OmniPaper to new languages such as Czech or Slovak. On the other hand, the MIRACLE research

[1] E-Content project Ref.: 22225, www.nedine.org.

team is applying new approaches to the image retrieval problem to take part in the next ImageCLEF workshop (CLEF 2005 edition at the time of writing). These approaches will try, among other approaches, to introduce a disambiguation step into the synonym query term expansion method applied. New languages are also being studied, especially Asiatic languages such as Japanese, Chinese, and Korean. For this purpose, the MIRACLE research team is taking part in the NTCIR[m] initiative.

Acknowledgments

This work has been partially supported by the projects OmniPaper (European Union, 5th Framework Programme for Research and Technological Development, IST-2001-32174), NEDINE (E-Content project Ref.: 22225), and GPS Project—Software Process Management Platform: modeling, reuse, and measurement (National Research Plan, TIN2004-07083). Special mention must be made to our colleagues, also members of the MIRACLE group: José C. González-Cristóbal, José M. Goñi-Menoyo, Julio Villena-Román, César de Pablo Sánchez, Josefa Z. Hernández Diego, and Santiago González Tortosa.

References

1. Baeza-Yates R, Riveiro-Nieto B. Modern Information Retrieval. Reading, MA: Addison Wesley; 1999.
2. Sparck Jones K, Willet P. Readings in information retrieval. San Francisco, CA: Morgan Kaufmann Publishers, Inc.; 1997.
3. Crestani F, Lalmas M, van Rijsbergen CJ, editors. Information retrieval: Uncertainty and logics. Norwell, MA: Kluwer Academic Publishers; 1998.
4. Peters C. Introduction. In: Peters C, editor. Workshop on Cross-Language Information Retrieval and Evaluation, CLEF 2000, Lecture Notes in Computer Science vol 2069. Berlin: Springer; 2001. pp 1–6.
5. Porter M. The Porter stemming algorithm. Available at: http://www.tartarus.org/~martin/PorterStemmer/, last accessed Nov. 30, 2004.
6. Snowball stemmers and resources. Available at: http://www.snowball.tartarus.org, last accessed Nov. 30, 2004.
7. Airio E, Keskustalo H, Hedlund T, Pirkola A. Multilingual experiments of UTA at CLEF 2003. The impact of different merging strategies and word normalizing tools. In: Peters C, Gonzalo J, Braschler M, Kluck M. Comparative evaluation of multilingual information access systems: Fourth workshop of the cross-language evaluation forum, CLEF 2003, Trondheim, Norway, August 21–22, 2003. Lecture Notes in Computer Science vol 3237. Berlin: Springer; 2004. pp 74–84.
8. López-Ostenero F, Gonzalo J, Verdejo F. Búsqueda de información multilingüe: estado del arte. Revista Iberoamericana de Inteligencia Artificial, No. 22. Asociación Española de Inteligencia Artificial; 2003. pp 11–25. Available at: http://www.aepia.org.
9. McNamee P, Mayfield J. JHU/APL experiments in tokenization and non-word translation. In: Peters C, Gonzalo J, Braschler M, Kluck M. Comparative evaluation of multilingual information access systems: Fourth workshop of the cross-language evaluation forum, CLEF 2003, Trondheim, Norway, August 21–22, 2003. Lecture Notes in Computer Science vol 3237. Berlin: Springer; 2004. pp 85–97.
10. Miller G. WordNet: A lexical database. Commun ACM 1995;38:39–41.

[m] NII-NACSIS Test Collection for IR Systems, http://research.nii.ac.jp/~ntcadm/index-en.html.

11. Vossen P, Bloksma L, Rodriguez H, Climent S, Calzolari N, Roventini A, Bertagna F, Alonge A, Peters W. The EuroWordNet base concepts and top ontology. Version 2. EuroWordNet (LE 4003) Deliverable, 1998.
12. García-Serrano A, Martínez P. An interface agent with linguistic skills. In: Moreno AM, van de Riet RP, editors. Applications of Natural Language to Information Systems (NLDB 2001). Lecture Notes in Informatics. Berlin: Springer; 2001. pp 45–54.
13. Berry M, editor. Survey of text mining. Clustering, classification and retrieval. Berlin: Springer Verlag; 2003.
14. Martínez-Santiago F. El Problema de la fusión de colecciones en la recuperación de información multilingüe y distribuida: Cálculo de la relevancia documental en dos pasos. Ph.D. Thesis, Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (UNED), 2004.
15. Cancedda N, Déjean H, Gaussier É, Renders JM, Vinokourov A. Report on CLEF-2003 experiments: Two ways of extracting multilingual resources from corpora. In: Peters C, Gonzalo J, Braschler M, Kluck M. Comparative evaluation of multilingual information access systems: Fourth workshop of the cross-language evaluation forum, CLEF 2003, Trondheim, Norway, August 21–22, 2003. Lecture Notes in Computer Science vol 3237. Berlin: Springer; 2004. pp 98–107.
16. Clough P, Sanderson M. The CLEF 2003 cross language image retrieval track. In: Peters C, Gonzalo J, Braschler M, Kluck M. Comparative evaluation of multilingual information access systems: Fourth workshop of the cross-language evaluation forum, CLEF 2003, Trondheim, Norway, August 21–22, 2003. Lecture Notes in Computer Science vol 3237. Berlin: Springer; 2004. pp 581–593.
17. Brill E. Some Advances in transformation based part of speech tagging. In: Proc 12th National Conference on Artificial Intelligence; 1994.
18. The GNU image-finding tool GIFT 0.1.9. Available at: http://www.gnu.org/software/gift/, last accessed Nov. 30, 2004.
19. Martínez-Fernández JL, Villena J, Fombella J, Serrano AG, Martínez P, Goñi JM, González JC. MIRACLE approaches to multilingual information retrieval: A baseline for future research. In: Peters C, Gonzalo J, Braschler M, Kluck M. Comparative evaluation of multilingual information access systems: Fourth workshop of the cross-language evaluation forum, CLEF 2003, Trondheim, Norway, August 21–22, 2003. Lecture Notes in Computer Science vol 3237. Berlin: Springer; 2004. pp 210–219.
20. Villena J, Martínez-Fernández JL, Fombella J, Serrano AG, Ruiz A, Martínez P, Goñi JM, González JC. Image retrieval: The MIRACLE approach. In: Peters C, Gonzalo J, Braschler M, Kluck M. Comparative evaluation of multilingual information access systems: Fourth workshop of the cross-language evaluation forum, CLEF 2003, Trondheim, Norway, August 21–22, 2003. Lecture Notes in Computer Science vol 3237. Berlin: Springer; 2004. pp 621–631.
21. Goñi JM, González JC, Moreno A. ARIES: A lexical platform for engineering Spanish processing tools. Nat Lang Eng 1997;3:317–345.
22. Goñi-Menoyo JM, González JC, Martínez-Fernández JL, Villena-Román J, García-Serrano AM, Martínez-Fernández P, de Pablo-Sánchez C, Alonso-Sánchez J. “MIRACLE’s hybrid approaches to bilingual and monolingual information retrieval. In: CLEF 2004 Workshop, Working Notes, Bath, UK, September 15–17, 2004.
23. Martínez-Fernández JL, Serrano AG, Villena J, Méndez Sáez VD, González Tortosa S, Castagnone M, Alonso J. MIRACLE at ImageCLEF 2004. In: CLEF 2004 Workshop, Working Notes, Bath, UK, September 15–17, 2004.
24. Cunningham H. GATE, a general architecture for text engineering. Comput Humanit 2002;36:223–254.