RIST

Information Processing at the Digital Age

VOL22. N.2(2018) 66-83

A pivot language based approach to multilingual document representation and information retrieval including Arabic

Souhila Boucham

Computer Science Department, USTHB University, Algeria
sbouchem@yahoo.fr

Submitted 28/01/2019, accepted 29/03/2019

Abstract

Arabic language has become an increasing interest in the field of Multilingual Information Retrieval (MIR). We deal in this work with the problem of Information Retrieval in a trilingual containing corpus documents in Arabic, French and English languages. We propose a language independent approach based on a pivot language.
The proposed approach combines a surface analysis and the Latent Semantic Analysis (LSA) statistical algorithm in a new way to break the terms of LSA down into units which correspond more closely to morphemes. These morphemes are the variable length character n-gram candidates extracted from different fragments separated by borders.
The obtained results are encouraging and competitive with state of the art results in multilingual field.

Keywords: multilingual document representation, multilingual information retrieval including Arabic, virtual document, principle of border, fragments and variable length character n-grams, parallel corpus, surface analysis and the LSA statistical algorithm, concept types, pivot language.

1. Introduction

Arabic language is one of the most widely spoken languages. This language has a complex morphological structure and is considered as one of the most prolific languages in terms of linguistic article. Therefore, Arabic Information Retrieval models need specific techniques to deal with this complex morphological structure (Emad et al., 2015).
Several research projects are investigating and exploring the techniques in monolingual or multilingual Information Retrieval systems for the English and European languages such as French, German, and Spanish and in Asian languages such as Chinese and Japanese. However, in Arabic language, there is little ongoing research in Arabic Information Retrieval systems or multilingual Information Retrieval systems including Arabic.

In a MIR system, a user expresses a query in his language and retrieves all relevant documents not only in his language but also in other languages.
The Core of a MIR system is the indexing process and the retrieval model : these models have to determine how documents and queries are represented, and how similarities between them are defined. In this study, we will focus on models which use indexing process to store data (documents and queries), and theories to compare documents representation versus queries representation.
A multilingual information retrieval system has also to address the problem of documents content representation and the of relevance evaluation problem. This evaluation is more difficult than in monolingual IR. Indeed, it is difficult to build a correspondence function with different languages of the documents and the query (Rekha and Sharvari, 2015).

To treat the language problem, MIR systems use as a basis for the indexing process different approaches such as document translation and query translation. Considering the limitations of these methods, other approaches propose to address multilingual indexing using a pivot language. The documents and queries are translated into the pivot language.

The purpose of this work is to explore the contribution of existing approaches to this problem in particular the use of resulting vectors of LSA model. Inspired by the idea proposed by (Roussey, 2001) which uses the **concept type** notion to represent the component of indexing language to improve the description of the documents in a multilingual context.
A text document is represented by a vector, where each dimension corresponds to a given **concept type** and where each value encodes the concept type importance in the document. LSA consists in "projecting" documents on a set of topics learned in an unsupervised way.

When processing a large corpus with a statistical tool, the first step typically consists of subdividing the text into information units called tokens. These tokens usually correspond to words, at least for the most part of them—"non words" tokens could be pictures, numbers, special characters or symbols. This tokenization process may appear to be quite simple, not to say trivial—tokenization, morphological analysis, and lexicons are discussed in (Pham et al., 2008), in the context of corpus-based processing. However, from an automated processing point of view, the implementation of this process constitutes a challenge. Indeed, how to reliably recognize words? What are the unambiguous formal surface markers that can delineate words, i.e. their boundaries? These questions are relatively easy to answer for languages such as French or English: basically, any string of characters delimited by a beginning space and an ending space is a simple word. But for many other languages, such as Arabic, the answer is much more complicated. In Arabic, subject pronouns and complements are sometimes attached to the verb. In this case, a token like katabtuhu " كتبته"corresponds in fact to a sentence (here, "I wrote it" or "I've written it") ("je l'ai écrit" in French). Obviously, the simple notion of tokens defined as strings of characters separated by spaces is an oversimplification that is highly inadequate for many situations and languages.

Considering the above, what then could constitute a reasonable atomic unit of information for the segmentation of a text, independently of the specific language it is written in?

we aim to overcome the language barrier by representing each document in the multilingual corpus by a set of concept types. The concept types comprise a pivot language for information representation and are defined in a semantic space representing a parallel corpus. It is an approach which combines surface analysis and statistical technique LSA for concept types detection. It combines techniques in a novel way to represent the terms of the input matrix of LSI down into units which correspond more closely to morphemes (variable length character n-grams candidates).

The paper is organized as follows : section 2 describes related work and approaches to multilingual information retrieval. The following section describes multilingual information retrieval including Arabic and then we elaborate our pivot language based approach to multilingual document representation in Section 4.
Section 5 presents the experimental results, and we evaluate the performance of the model in Section 6. Discuss the evaluation of our model in Section 7 before concluding.

2. Approaches to multilingual information retrieval

2.1. Document based representation approaches

Several approaches have been proposed to solve the MIR problem.

In (Susan et al., 1996), the authors adapt a CL-LSI method to Cross Language (CL) retrieval.
The Basic idea consists in considering an initial sample of documents translated by human or, perhaps, by machine, to create a set of dual-language training documents. The LSI method ignores word order and, therefore, treats this document as a bag of freely intermingled French and English words.
Such a set of training documents is analyzed using LSI, and the result is a semantic space with a reduced dimension in which related terms are close to each other.
Because the training documents contained both French and English terms, the LSI space will contain terms in both languages.
The next step in the CL-LSI method is to add documents in just French or English. This is done by locating a new document at the weighted vector sum of its constituents terms. The result of this process is that each document in the database, whether it is in French or in English, has a language- independent representation in

terms of numerical vectors. Users can now pose queries in either French or English and get back the most similar documents regardless the language.

In (McNamee and Mayfield, 2004), the authors proposed an n- gram based approach to IR for NLP applications, the n-grams of characters have a constant number of characters defined.

All n-grams were of the same length. This approach is applied specifically for *CLIR Systems.*
However, the results of CLIR are exclusively for European languages written into the Latin alphabet. This is why they are obtaining in their works 'surprisingly good results without translation of the request' and without the help of LSI in any form.

HYBRED (HYBrid Representation of Documents), is a text classification approach for optical character recognition documents. This approach combines different features in a single relevant representation (Sami et al., 2009). The principle of this approach is summed up in the following steps:
1. Selection according to a Part-of-Speech tag:
Select only the terms belonging to one or more grammatical categories such as nouns and verbs.

2. Application of the principle of border.

The LEXTER system (Bourigault, 1994), proposed by D. Bourigault is a terminology extraction tool. This system extracts maximum nominal phrases by identifying border marker. These boundaries are "preposition + possessive adjective", "preposition + determiner", etc.). The candidate terms are extracted by the use of the border marker information.
In (Sami et al., 2009) study, the words giving less information are replaced by a border. The objective is the same as the LEXTER system. The difference is that in (Sami et al., 2009) borders are the words/tags less relevant to the classification task (adverb, preposition, etc.).

*3.*The next step is to represent the characters with N-grams. It is a merger of N-grams of different fragments separated by the border.

4. a filter based on statistical TF.IDF is used to assign weights to features.

Within the framework of information retrieval, one interest is to select relevant documents related to a user request among a predefined collection of documents. In this context, defining a “similarity measure” between documents and users’ requests is a central issue that shapes the core component of information retrieval systems (IRS) namely the comparison engine. This measure also highly influences the performance of IRS. An approache presented in (Yaël, 2009) where the document/request comparison consists in comparing document and request on the basis of their relation system. This approach exploits the information conveyed by the relationships between terms and documents, between terms and between documents.
This approach follows (Glen and Jennifer, 2002) where the authors motivated by mining the space of graph properties, propose a general method for measuring the structural similarity of a graph and then for studying the resemblance between the nodes which compose this graph.

We survey other approaches based on N-grams but now in monolingual document representation : N-gram based on low-dimensional representation for document classification (Rémi and Ronan, 2015).
The model is divided into three steps:

1) vector representations of n-grams are obtained by averaging pre-trained representations of its individual words. From word vector representations obtained with the Skip-gram model (Mikolov et al., 2013) or GloVe (Pennington et al., 2014) models, n-grams with different length n can then be embedded in the same dimensional vector space with a simple element-wise addition. This makes it possible to compute distances between n-grams.
2) n-grams are gouped into K semantic concepts by performing K-means clustering on all n-gram representations;
3) documents are represented by a bag of K semantic concepts, where each entry depends on the presence of n-grams from the concepts defined in the previous step.

2.2. word based representation approaches

We have previously discussed in a previous sub-section techniques directly related to our work. In this sub-section, we survey other approaches in multilingual distributed representations and training bilingual word representations.
Distributed meaning representations are a natural way to encode covariance relationships between words and phrases in naturel langage procecing (NLP).
Following the idea that human language acquisition is widely seen as grounded in sensory-motor experience (Paul, 2001; D. Roy, 2003; Karl Moritz and Phil, 2014a), there have been some attempts to use multi-modal data for learning better vector representations of words. Such methods, however, are not easily scalable across languages or to large amounts of data for which no secondary or tertiary representation might exist.
(Karl Moritz and Phil, 2014a) has proposed a novel method for inducing cross-lingual distributed representations for compositional semantics. It is a novel method for learning vector representations at the word level and beyond.
This model learning assign similar embeddings to aligned sentences and dissimilar ones to sentence which are not aligned while not requiring word alignments. It attempts to learn semantics from multilingual data, in condition that must exists enough parallel data, a shared representation would be forced to capture the common elements between sentences from different languages. What two parallel sentences have in common, of course, is the semantics of those two sentences.
However, these requirements reduced the chance to work with low-resource languages such as multilingual resource including Arabic.

The author has evaluated his models using the cross-lingual document classification (CLDC) task. The CLDC experiment focused on establishing the semantic content of the sentence level representations, also briefly investigate the induced word embeddings.

Most of the recent works in bilingual representation learning such as (Karl Moritz and Phil, 2014b; Tomas et al., 2014; Stephan et al., 2014) only focus on learning embeddings for words and use simple functions to synthesize representations for larger text sequences from their word members. In contrast, (Hieu et al., 2015) aims to learn representations for phrases and sentences as a whole so as to represent non-compositional meanings.
The work of (Karl Moritz and Phil, 2014b; Sarath et al., 2014) are similar to that of (Hieu et al., 2015) in eliminating the monolingual components and just training a model with bilingual objective to pull distributed representations of parallel sentences together.

These first two approaches, however, only use simple bag-of-word models to compute sentence representations and has a potential disadvantage incapturing the non-compositional meanings of sentences. Instead, (Hieu et al., 2015) learn representations for text sequences as a whole, but in the bilingual context.

Recent multilingual applications in training bilingual representations where similar-meaning words in two languages are embedded close together in the same highdimensional space. However, most bilingual representation works tend to focus on learning embeddings that are tailored towards achieving good performance on a bilingual task, often the crosslingual document classification task, but it doesn't preserve clustering structures of monolingual word representations.

In the work of (Minh-Thang et al., 2015), such a goal of learning representations of high quality both bilingually and monolingually was achieved through a joint learning approach. Specifically, the joint model utilizes both the context co-occurrence information present in the monolingual data and the meaning equivalent signals exhibited in the parallel data.

Finally, while words in documents are generally treated as discrete entities, they can be embedded in an Euclidean space which reflects a priori notion of similarity between them. In such a case, a text document can be viewed as a bag-of embedded-words : a set of real valued vectors.

(Stéphane and Florent, 2013) proposed a novel document representation based on a continuous word embeddings. It consists in non-linearly mapping the word embeddings in a higher-dimensional space and in aggregating them into a document level representation. However, in this work the author questions the treatment of monolingual documents. Indeed, intuitvely some words are closer to each other from a semantic stand point.

3. Multilingual information retrieval including Arabic

(Aliane, 2006) described an approach to indexing and retrieval of documents in : Arabic, French and English. The proposed system is founded on a knowledge representation formalism introduced by (Roussey, 2001),

namely semantic graphs which supports a domain ontology. Documents and queries are also represented in this formalism.
the aim of this work is to develop an ontology which supports indexing, retrieval and information extraction. the design and development of this system proceeds in two steps: the first step builds in partnership with a human expert a semantic graph based ontology which explicit the domain's concepts and their relationships. In a second step, the ontology is considered as a bootstrap that initializes the system's knowledge and thus the indexing process is based on a linguistic method, precisely, repeated segments extraction algorithm. When identified, repeated segments are submitted to a filtering procedure using linguistic filters.
The retrieval system consists of a graph comparison to find relevant documents for the extended user query.

In (Peter et al., 2008), the authors described an entirely statistics-based, unsupervised, and language independent approach to MIR, this approach has an important theoretical advantage over LSI: it combines well-known techniques in a novel way to break the terms of LSI down into units which correspond more closely to morphemes. Thus, it has a particular appeal for use with morphologically complex languages such as Arabic.
Since the approach of n-gram tokenization has the advantages of being entirely statistically- based and language-independent, the authors examined whether it could be combined with LSI to allow CLIR.

The set of experiments was guided by the intuition that not all n-grams are morphologically significant.
The weighting for one token has to be contingent on the weighting for another in the same term. An alternative is to select the tokenization which maximizes mutual information (MI). The pointwise MI of a pair s_1 and s_2 as adjacent symbols is:

$$MI = \log P(s_1\, s_2) - \log P(s_1) - \log P(s_2)$$

If $s1$ follows $s2$ less often than expected on the basis of their independent frequencies, then MI is negative; otherwise, it is positive.
To obtain MI, it needs to compute the log probability (logp) of every n-gram in the corpus.
If S_k ($k = 1, \ldots, K$) denotes the set of all ngrams of length k, and s_n is a particular n-gram of length n, then the compute logp for s_n as:

$$logp = \log F(s_n) - \log \sum (F(S_n))$$

where $F(s_n)$ is the frequency of s_n in the corpus, and $\sum(F(S_n))$ is the sum of the frequencies of all S_n in the corpus.
In all cases, logp is negative, and MI is maximized when the magnitude of the sum of logp for all elements in the tokenization (also negative) is minimized, i.e. closest to zero.
Tokenizations consisting of one, two or more elements (respective examples are comingle, co+mingle, and co+ming+le) will all receive a score, although those with fewer elements will tend to be favored.
Guided by (McNamee and Mayfield, 2004) finding that there is an optimal (language-dependent) value of k for S_k, the authors varied the maximum length of n-grams allowed in tokenizations.
finally, the authors have demonstrated LMSA, a linguistically (specifically, morphologically) more sophisticated alternative to LSI. By computing mutual information of character n-grams of non-fixed length.

4. A pivot language based approach to multilingual document representation

Some methods propose to cross the language barrier using a pivot language. This language is used to represent the document and the query regardless of the source and target languages. All the problem then is the definition of pivot language for multilingual information retrieval and conversion and enconversion between natural language and this language.
Inspired from the idea proposed by (Roussey, 2001) which uses **concept types** to improve the description of documents in a multilingual context to represent the components of our pivot language.

A set of concept types corresponds to a domain conceptualization. The process of conceptualization is:

- Identify the domain or documents ie define the domain by identifying what type of documents will be indexed using this modelization. As stated in (Roussey, 2001), the corpus delimits the domain.

- Identify the concept types. That is to say represent each concept by an identifier which is not a term, because several terms may designate it and doesn't depend only on the language of the document. Therefore, these concept types represent viewpoints on the domain objects.

4.1. A semantic space as pivot language

In concept classification, the objects considered as similar are grouped in the same group.

Generally, the work in constructing multilingual semantic-space models divides into two main streams: (a) those that make use of comparable or parallel corpora and (b) those that only require unaligned or monolingual text. The former includes various extensions to standard techniques such as bilingual latent semantic models (LSA) (Yik-Cheung and Tanja, 2007; Nick and Marcello, 2011) or bilingual/multilingual topic models (LDA) (Bing and Eric, 2007; XiaochuanNi et al., 2009; David et al., 2009; Ivan et al., 2011).

We adapt the bilingual latent semantic models in our multilingual approach. The general assumption is that aligned documents share identical topic distributions.

The approach of LSA (Dumais et al., 1988) is one of the most successful semantic memory models. It is founded on the calculation of co-occurrences in vector from a corpus. A calculation of singular vectors and size reduction are performed so as of the one part apply a form of transitivity and partly reduce noise.

LSI is an unsupervised approache and an automatic statistical method for cross-language information retrieval that re-describes the textual data in a new smaller semantic space.

LSI uses a Matrix Computation as a basis. All other strategies directly matche key words. Since the same concept can be described using many different keywords, this type of matching is prone to failure. People used the same query for same concept only 20% of time. The goal is to represent the underlying semantics of documents rather than matching keywords. LSI uses Singular Value Decomposition to capture this semantic structure. This filters noise found in a document, such that two documents that have the same semantic concepts are located close to one-another in a multi-dimensional space.
we focus on the representation of the thematic (not a semantic) aspects of textual segments such as documents by concept_type *document vectors.
Thematic analysis will be here seen as the calculation of a vector allowing to represent the lexical fields of a document of our parallel corpus.

There are three major advantages of using the LSI representation with the following labels: synonymy, polysemy and term dependency.

Synonymy: Synonymy refers to the fact that the same underlying concept can be described using different terms. Traditional retrieval strategies have trouble discovering documents on the same topic that use a different vocabulary. In LSI, the concept in question as well as all documents that are related to are all likely to be represented by a similar weighted combination of indexing variables.

Polysemy: Polysemy describes words that have more than one meaning, which is a common property of any language. Large numbers of polysemous words in the query can reduce the precision of a search significantly. By using a reduced representation in LSI, one hopes to remove some "noise" from the data, which could be described as rare and less important usages of certain terms. (however, this would work only when the real meaning is close to the average meaning. Since the LSI term vector is just a weighted average of the different meanings of the term, when the real meaning differs from the average meaning, LSI may actually reduce the quality of the search).

Term Dependency: The traditional vector space model assumes term independence. Terms serve as the orthogonal basis vectors of the vector space. Since there are strong associations between terms in a language, this assumption is never satisfied. While term independence represents the most reasonable first-order approximation, it should be possible to obtain improved performance by using term associations in the retrieval process. Adding common phrases as search items is a simple application of this approach. On the other hand, the LSI factors are orthogonal by definition, and terms are positioned in the reduced space in a way that reflects the correlations in their use across documents. It is very difficult to take advantage of term associations without dramatically increasing the computational requirements of the retrieval problem. While the LSI solution is difficult to compute for large collections, it needs only to be constructed once for the entire collection and the performance at retrieval time is not affected.

The fundamental aim of an LSI model is to achieve a conceptual representation of documents. The LSI technique applied to MIR may be seen as the introduction of a pivot language by changing the

expression space of index vectors on new dimensions concretizing the pivot: the document and the query are represented in a common space independent from a language. This approach is based on the vector model. The whole problem lies in defining the vector space.

4.2. Definition of the vector space

The vector formalization of a document - which reduces it to an unordered list of index terms - sufficient to reveal resemblances, semantic proximities between documents (docs / queries) in a corpus. The problem posed is to find, regardless of the language or script, the descriptors or basic unit of information that are identifiable and extractable well as most relevant to a document collection written in a given language.

(Balpe et al., 1996) suggest that this unit should be defined according to the goal we set ourselves when reading or processing a text. More precisely, from a numerical classification based knowledge extraction viewpoint, the definition of the basic unit of information to be considered depends on the following:

- The unit of information must be a portion of the input text submitted to the numerical analysis processor;
- From an automated processing point of view, it should be easy to recognize these units of information;
- The definition of the unit of information should be independent of the specific language the text is written in;
- The units of information must be statistically meaningful when evaluated or compared between themselves. It should be easy to compute their frequencies in various parts of the input text, as well as to estimate their distribution and the regularity with which some units co-occur in certain portions(segments) of the text.

In the field of textual representation of data the state of the art consists of three approaches:

1) An approach based on the notion of characters string - a string represents an elemental unit of sense delimited by trivial separators, such as the spaces or punctuation signs,

2) An approach based on the concept of word defined in cases as an inflected form lemmatized, term or multi-term.

3) An approach based on the notion of N-gram, an N-gram is a sequence of N consecutive characters.

Although, it makes tokenization (where words are considered as tokens) relatively easy in English or French, it is much more difficult for other languages such as Arabic. Moreover, stemming or lemmatisation, typically used to normalize and reduce the size of the lexicon, constitutes another challenge. The notion of n-grams which, for the last ten years, seems to have produced good results both in language identification and speech analysis, has recently become a privileged research axis in several areas of knowledge acquisition and extraction from text.
The concept of n-grams was first discussed in 1951 by Shannon (Shannon, 1950). Since then the concept of n-grams has been used in many areas, such as spelling-related applications, string searching, prediction and speech recognition.

A character N-gram is a set of n consecutive characters extracted from a text.
An n-gram is a character sequence of length n extracted from a document. Typically, n is fixed for a particular corpus of documents. To generate the n-gram vector for a document, a window of n characters in length is moved through the text, sliding forward one character at a time. At each position of the window, the sequence of characters in it is recorded. For example, the first four 5-grams in the sentence " character sequences..." are " char", "chara", "harac" and "aract". In some schemes, the window may be slided more than one character after each n-gram is recorded.

- *Advantages of n-grams encoding*

There are several advantages for using n-grams:

First, the system can be garble tolerant by using n-gram as a basic term. If a document is scanned using OCR (Optical Character Recognition), there may be some misread characters. For example, suppose "character" is scanned as "claracter". The word-based system will not be able to match this word because it is misspelled, but an n-gram based system will still match the other n-grams such as "aract", "racte"… and take their frequency into account. From this, we can see that by using n-gram technology the system can be garble tolerant.

Second, by using n-grams the system can achieve language independence. In a word-based information retrieval system there is language dependency. For example, in some Asian languages, different words are not separated by spaces, so a sentence is composed of many consecutive characters. Grammar knowledge is needed to separate those characters into words, which is a very difficult task to perform. Using n-grams, the system does not need to separate characters into words.

Additionally, n-gram based systems do not use stop words. This is because the number of unique n-grams in a document is very big and distribution is very wide. There are few n-grams that have high frequency. From Ekmekcioglu's research (Cuna et al., 1996), stop words and stemming are superior for word-based system but not significant for an n-gram based system.

4.3. Creation of the semantic space as a pivot language

To create the semantic space the development of our approach proceeds in two phases:

A. Corpus building : In this first phase, the constitution of our trilingual corpus (Arabic, French and English).

B. The second phase comprises the following successive steps:

1. Our approach consists in applying a preprocessing step on the selected textual document issued from the corpus. Then, we apply the process of feature extraction. This approach consists not only in extracting the most relevant words of sentences issued from a textual document but also we attempt to reduce the dimensionality of the corpus by removing the non- informative words.

2. Take the documents of the three languages, concatenated to create a set of virtual documents (a set of dual-language training documents). The virtual document is the concatenation of source document + its translations in two target languages).

3. Analysis stepe: the virtual document is considered as one document regardless of the language. The set of documents is analyzed by multilingual LSI. The terms of the term-document matrix are the morphemes results of the preprocessing step from the corpus (character n-grams candidates of non-fixed length). LSI examines the similarity of the contexts in which terms appear, and creates a reduced-dimension feature-space representation where terms that occur in similar contexts are near each other.

4. The result is a reduced semantic space that will serve as a pivot language where related terms are grouped in the same concept. Thus, the concepts comprise a pivot language for information representation and are defined in a semantic space representing a parallel corpus.

5. The next step is to represent the documents in each language around the space terms.

The global architecture of our approach is as follows:

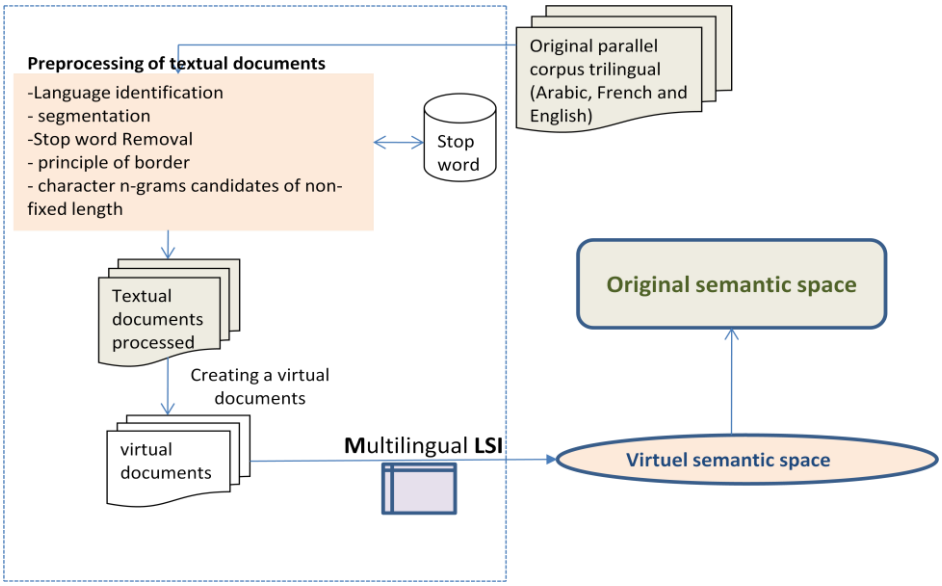

Fig. 1. Approach architecture

The preprocessing step is thoroughly detailed in the following:

4.3.1. Segmentation

Firstly, we apply the process of Case Folding which aims to reduce all letters to lowercase (i.e., in case of English texts), (nevertheless, some exceptions have to be considered). A textual document is not just a sequence of words, but it has a coherent structure. The meaning of each word cannot be determined until it is placed in the structure of the text. Recognizing the structure of the text is of high importance. One of the most important constituents of the text structure is the text segment. A text segment, whether or not it is explicitly marked, i.e. sentences and paragraphs, is defined as a sequence of sentences that display local coherence. Hence, the segmentation consists in determining the relative positions of each text stream which ended by a full stop. In order to process and understand a document, a text segmentation phase is useful. This phase consists in converting text into lower-cases, remove punctuation and recognize the pattern numbers, i.e. 10.5, and rewriting them in a more meaningful pattern. The paragraph breaks including the end of the sentences should be kept. Nevertheless, some difficulties may occur in the segmentation step:
- Removal of any sequences of successive dots or ellipsis.
- The special characters like ?, ! or any other special character, sometimes indicate the end of a sentence and it needs a particular processing.
At the end of this step, we generate a list of segments. Each segment is marked by a starting offset (the beginning of the segment) and an ending one (the end of the segment). These offsets allow referencing any extracted segment.

4.3.2. Stop word Removal

During the indexing phase, we can determine, a priori, the significant words of each document because each word within the document is a potential keyword.
However, we may eliminate some words like those which are too frequent. In (Radwan and Jean-Hugues, 2002), there is the principle of filtering out parasite words. In practice, the method suggested yields a long list of words among which a number of parasite words, that is words, though otherwise uninteresting, happen to contain one of the n-grams characteristic.
In this phase, we proceed, for each selected segment, by the removal of the stop words which have meaningless information. This will save around 25% to 30%. In addition, the corpus should be exempted from empty words (i.e., or, and, of, it, we, is, are and so on.). These empty words are grouped as common words and recorded in a database table. They should be removed from the document.
Empty words can be eliminated using a predefined list (stop list) for French and English. For French, this list, for example, contains mainly articles, pronouns, few adverbs as {"a", "alors", "après", ect}.
Arabic is a morphologically rich language with a large set of morphological features such as person, number, gender, voice, aspect, case, and state. Arabic features are realized using both concatenative (affixes and stems) and templatic (root and patterns) morphology with a variety of morphological, phonological and spelling

adjustments. In addition, Arabic has a set of very common clitics that are written attached to the word, e.g., the conjunction و+ w+ 'and', the preposition + ب b+2 'with/in', the definite article + ال Al+ 'the' and a range of pronominal clitics that can attach to nouns (as possessives) or verbs and prepositions (as objects).
Therefore, Arabic is an agglutinative language. Articles, prepositions and pronouns stick to adjectives, nouns, verbs and particles to which they relate, which generates ambiguity in morphological analysis of words. So all these empty words of Arabic can be concatenated together.

For example: for "تلك" we can derive 'فبتلك' = ف+ب+تلك et.وتلك = و+تلك
The list of stop words for Arabic includes particles (الحروف), pronouns (الضمائر), demonstratives (أسماء الاشارة), conditional word (اسماء الشرط), etc..

4.3.3. N-grams of characters

The representation of data with the N-grams of characters is motivated by the data complexity (multilingual data).
We conducted a search of N-grams regardless of their size. This choice is justified by the fact that it frees from the notion of word, so from any morphosyntactic analysis.
In the next example, we introduce our approach.
We consider the sentence " le bijoux plaqué or a du charme (the goldplated jewellery has charm)".

- The selection and removal of stop words returns the following results: "bijoux plaqué or a charme".

After this initial processing, we represent the words from the N-grams of characters. The application of N-grams of characters process gives three possibilities of representation:

1. The first representation is based on a bag of selected words. The application of N-grams where N = 5 gives the following result: « _bijo, bijou, ijoux, joux_, oux_p, ux_pl, x_pla, _plaq, plaqu, laqué, aqué_, qué_o, ué_or, é_or_, _or_a, or_a_, **r_a_c, _a_ch, a_cha**, _char, charm, harme, arme_ ».
This application is flawed because it adds noise and unnecessary N-grams. For example: "a_cha"' is a Ngrams which represents noise (N-grams from the fragment "a du charme (has charm)" where the word "du" was deleted). Indeed, the elimination of the stop words of the initial sentence returns irrelevant results.

2. A second kind of representation is based on the Ngrams of characters application for each extracted word separately. As result, we have: " _bijo, bijou, ijoux, joux_, _plaq, plaqu, laqué, aqué_, _char, charm, harme, arme_".
This representation corrects the defects caused by the previous method but provides fewer data (in particular with short words). For example, by using the N-grams characters with $N \geq 5$, the noun "or" cannot be identified. This deletion causes a loss of information.

4.3.4. Application of the principle of border

The two representations mentioned above have major defects with the introduction of **noise** (first method) and **silence** (second method). Thus, we have introduced a principle of border. In our study, the words giving less information (i.e. stop words list) are replaced by a border.
The objective is the same as the LEXTER system or the HYBRED approach. The difference is that our borders are list of stop words and do not rely on linguistic rules as in LEXTER or the words/tags less relevant to the classification task (adverb, preposition, etc.) as in HYBRED.
This method corrects the noise added during the first proposed treatment. it takes into account groups of words (e.g. "plaqué or"). The result according to the principle of border is shown below:
" X bijoux plaqué or a X charme", "X" represents the border.
Then we can extract the 5-grams of characters in the two fragments of the text (i.e. " bijoux plaqué or a " and " charme "): " _bijo, bijou, ijoux, joux_, oux_p, ux_pl, x_pla, _plaq, plaqu, laqué, aqué_, qué_o, ué_or, é_or_, _or_a, or_a_, _char, charm, harme, arme_ ". The proposed algorithm is organized as follows:

Inputs: The set of multilingual documents forming the corpus.
Outputs: Matrix.
For all documents do:
1. Segmentation and removal of stop words.
2. Application of the principle of border.
3. Representation of words extracted with the character N-grams of variable length.
4. Assigning weights based on statistical measure (entropy).
End.

In the following, we develop a complete example of the proposed approach. We consider the sentence "Il faut une infinie patience pour attendre toujours ce qui n'arrive jamais".

1. The Segmentation and removal of stop words list returns: "faut infinie patience attendre arrive".

2. The application of the principle of border, gives us: "X faut X infinie patience X attendre X arrive X".
3. The N-grams of characters representation, where N = 3 returns:

Word	**N-grams of characters**
[_faut_]	[_fa, fau, aut, ut_]
[_infinie patience_]	[_in, inf, nfi, fin, ini, nie, ie_, e_p, _pa, pat, ati, tie, ien, enc, nce, ce_]
[_attendre_]	[_at, att, tte, ten, end, ndr, dre, re_]
[_arrive_]	[_ar, arr, rri, riv, ive, ve_]

Thus, we can calculate the sum of all 3-grams: N-grams("_faut_") + N-grams("_infinie patience_") + N-grams("_attendre_") + N-grams("_arrive_").

we obtain : {_fa, fau, aut, ut_, _in, inf, nfi, fin, ini, nie, ie_, e_p, _pa, pat, ati, tie, ien, enc, nce, ce_, _at, att, tte, ten, end, ndr, dre, re_, _ar, arr, rri, riv, ive, ve_}

Finally, digital filtering will be applied. This filter is used to reduce the number of features in step 3 (N-grams representation).
After applying these different stages we obtain a representation for each text of the corpus.

The length n of n-grams is not fixed since every language has its own properties.
one in which all n-grams were of the same length (as per (McNamee and Mayfield, 2004)), and one in which n-grams of different lengths were mixed.

As in (Peter et al., 2008) we propose a character selection method of non-fixed length based on repeated n-grams that is already filtered by the principle of border.
After extracting of all candidate tokens for each final n-gram, we filter these candidates to select the single candidate which best represents the final n-gram and maximizes mutual information (MI).

$$IM_{max}(\text{n-gram}) = \max_{i=1}^{n} \{IM(S_i)\}$$

S_i (i = 1, …, n) denotes the set of all n-grams of length i. for example:
For "_faut_" : 'f+a+u+t', 'fa+u+t', 'f+au+t', 'f+a+ut', 'fa+ut', 'fau+t', 'f+aut', 'faut' are the candidates, but 'fau+te' and 'fau+aut' are not.

5. Experimental Evaluation

Even though Arabic is a language which is spoken by over 422 million people in over 22 countries (Emad et al., 2015). The Quran, the holy book of Islam, gave the language a considerable geographic expansion. Automatic processing of Arabic is considered difficult to understand, because of morphological and structural features, such as polysemy, irregular forms of some words and its derivative and concatenative properties.
The initial objective was to provide a platform for the construction of a pivot language from multilingual texts (Arabic, French and English). To do this, it was necessary to have a parallel trilingual corpus (Arabic, French and English).

5.1. Building and results of the first corpus

The texts are extracted from websites http://www.lexilogos.com/coran.htm.
(http://www.alargam.com/quran2/quran3/index.htm in Arabic, the Quran in English, the Koran concordance tables & French).
In a first step, since the Quran comprises 114 Surats, the corpus consists of 114 documents in each language (ie 342 documents) and 100 types of queries verse, whose relevance is evaluated manually through web search. We will work on the corpora in a raw format. An example of a query is given in the following figures (2 and 4):

" الله لا إله إلا هو الحي القيوم لا تأخذه سنة و لا نوم له ما في السموات و ما في الأرض من ذا الذي يشفع عنده إلا بإذنه يعلم ما بين أيديهم و ما خلفهم و لا يحيطون بشيء من علمه إلابما شاء وسع كرسيه السموات و الأرض و لا يؤوده حفظهما و هو العلي العظيم" البقرة 255

Fig. 2. Query verse of al-Kursi number 255 of Surah the cow (Al-Baqarah)

The top 20 closest documents to this query are given in the following figure:

document title
1. سورة البقرة رقم 2 الله لا إله إلا هو الحي القيوم لا تأخذه سنة و لا نوم له ما في السموات و ما في الأرض من ذا الذي يشفع عنده إلا بإذنه يعلم ما بين أيديهم و ما خلفهم و لا يحيطون بشيء من علمه إلابما شاء وسع كرسيه السموات و الأرض و لا يؤوده حفظهما و هو العلي العظيم (البقرة 255 **)** "
2. سورة آل عمران رقم 3 الم (1) الله لا اله الا هو الحي القيوم (2)
3. سورة غافر رقم 40 حم * تَنزِيلُ الْكِتَابِ مِنَ اللَّهِ الْعَزِيزِ الْعَلِيمِ * غَافِرِ الذَّنْبِ وَقَابِلِ التَّوْبِ شَدِيدِ الْعِقَابِ ذِي الطَّوْلِ لا إِلَهَ إِلَّا هُوَ إِلَيْهِ الْمَصِيرُ [غافر:1-3]
4. سورة طه رقم 20 (4. و عنت الوجوه للحي القيوم و قد خاب من حمل ظلما (111
5. سورة الحاقة رقم 69 كَذَّبَتْ ثَمُودُ وَعَادٌ بِالْقَارِعَةِ)4(
6. **سورة الشورى رقم 42** فَلِذَلِكَ فَادْعُ وَاسْتَقِمْ كَمَا أُمِرْتَ وَلَا تَتَّبِعْ أَهْوَاءَهُمْ وَقُلْ آمَنْتُ بِمَا أَنْزَلَ اللَّهُ مِنْ كِتَابٍ وَأُمِرْتُ لِأَعْدِلَ بَيْنَكُمُ اللَّهُ رَبُّنَا وَرَبُّكُمْ لَنَا أَعْمَالُنَا وَلَكُمْ أَعْمَالُكُمْ لَا حُجَّةَ بَيْنَنَا وَبَيْنَكُمُ اللَّهُ يَجْمَعُ بَيْنَنَا وَإِلَيْهِ الْمَصِيرُ (15](لَهُ مَا فِي السَّمَاوَاتِ وَمَا فِي الْأَرْضِ وَهُوَ الْعَلِيُّ الْعَظِيمُ (4
.7 **سورة الواقعة رقم 56**
8. سورة النازعات رقم 79 (ءانتم اشد خلقا ام السماء بنيها (27) رفع سمكها فسويها (28) و اغطش ليلها و اخرج ضحيها (29) و الأرض بعد ذلك دحيها (30
9. سورة لقمان رقم 31 وْ أَنَّمَا فِي الْأَرْضِ مِن شَجَرَةٍ أَقْلَامٌ وَالْبَحْرُ يَمُدُّهُ مِن بَعْدِهِ سَبْعَةُ أَبْحُرٍ مَّا نَفِدَتْ كَلِمَاتُ اللَّهِ إِنَّ اللَّهَ عَزِيزٌ حَكِيم (27 الم - تِلْكَ آيَاتُ الْكِتَابِ الْحَكِيمِ - هُدًى وَرَحْمَةً لِلْمُحْسِنِينَ - الَّذِينَ يُقِيمُونَ الصَّلَاةَ وَيُؤْتُونَ الزَّكَاةَ وَهُمْ بِالْآخِرَةِ هُمْ يُوقِنُون .[4-3- 2 -1]
10. سورة الملك رقم 67
11. سورة القدر رقم 97
12. سورة التغبن رقم 64
13 . سورة الاحقاف رقم 46
14. سورة المؤمنون رقم 23
15. سورة الروم رقم 30 في سورة الروم (فَسُبْحَانَ اللَّهِ حِينَ تُمْسُونَ وَحِينَ تُصْبِحُونَ (17) وَلَهُ الْحَمْدُ فِي السَّمَاوَاتِ وَالْأَرْضِ وَعَشِيًّا وَحِينَ تُظْهِرُونَ (18))
16. سورة هود رقم 11 وَهُوَ الَّذِي خَلَقَ السَّمَاوَاتِ وَالْأَرْضَ فِي سِتَّةِ أَيَّامٍ وَكَانَ عَرْشُهُ عَلَى الْمَاءِ لِيَبْلُوَكُمْ أَيُّكُمْ أَحْسَنُ عَمَلًا وَلَئِنْ قُلْتَ إِنَّكُمْ مَبْعُوثُونَ مِنْ بَعْدِ الْمَوْتِ لَيَقُولَنَّ الَّذِينَ كَفَرُوا إِنْ هَذَا إِلَّا سِحْرٌ مُبِينٌ (7
..... .
26. سورة مريم رقم 19

Fig. 3. The 20 closest Documents

To see the judgment of the documents relevance returned by the search. A surat in the quran may be related to other surats. Our experimentation shows that the top results are those surats directly related to the query surat.
A second example of a query is given in figure below:

{ وَأَلَّفَ بَيْنَ قُلُوبِهِمْ لَوْ أَنفَقْتَ مَا فِي الأَرْضِ جَمِيعاً مَّا أَلَّفَتْ بَيْنَ قُلُوبِهِمْ وَلَـكِنَّ اللّهَ أَلَّفَ بَيْنَهُمْ إِنَّهُ عَزِيزٌ حَكِيمٌ } الأنفال 63

Fig. 4. verse 63 of Surah number Query booty (Al-Anfal)

The verse Number 63 of Surah booty (Al-Anfal) constitutes one of the verses of the magistrate (من آيات الصلح).

The first closest documents of this query are given in the following figure:

document title
1. سورة الأنفال { وَأَلَّفَ بَيْنَ قُلُوبِهِمْ لَوْ أَنفَقْتَ مَا فِي الأَرْضِ جَمِيعاً مَّا أَلَّفتْ بَيْنَ قُلُوبِهِمْ وَلَـكِنَّ اللّهَ أَلَّفَ بَيْنَهُمْ إِنَّهُ عَزِيزٌ حَكِيمٌ } (الأنفال 63)
2- سورة الممتحنة رقم 60 { عسى الله أن يجعل بينكم وبين الذين عاديتم منهم مودة ورحمة } (الممتحنة 7)
3. سورة الرعد رقم 13 (يَصِلُونَ مَا أَمَرَ اللَّهُ بِهِ أَنْ يُوصَلَ) سورة الرعد 13 فتحت ذلك صلة الأرحام والقرابات، وتحته الصلة القائمة بين الناس بسبب الإيمان، وذلك بالإحسان إليهم قدر الطاقة، ونصرتهم، والنصيحة لهم؛ فتلك أوثق عرى المحبة . انظر تفسير سورة الرعد في الكشاف للزمخشري.

Fig. 5. The first Documents closest

The same principle is applied to retrieve to a written question in one language (English or French) documents written in another language closest one.
The results obtained are encouraging and variability shows that they are perfectible.
Most queries are processed verses and deal with a specific topic.

5.2. Second test collection (CISI)

For the requirements of certain stages of our application, we needed to work on another corpus, because we need to compare the results in terms of relevant documents for a given query in our corpus in relation others, we used the corpus CISI for our tests. This corpus account 1460 documents and 112 queries.
The indexing of documents is based on the steps previously mentioned. Queries are considered and treated as documents.
After indexing, the corpus CISI account 38821 terms to 38821 for 1460 documents.
The CISI corpus has numerous queries that have no or few really deemed relevant documents. For this reason, we limited our study to all queries having at least 10 relevant documents ie 67 requests.

5.2.1. Evaluation criteria

We have chosen to compare our measure to the cosine measure and the SimRank measure (Yaël, 2009).
The original approach presented in (Yaël, 2009) consists in comparing document and request on the basis of their relation system. This approach exploits the information conveyed by the relationships between terms and documents, those between terms and those between documents.

In order to evaluate our model, we have used the MAP measure (Mean Average Precision), which widely accepted measure in the evaluation of the performance of information retrieval systems.
the average precision provides a global view of the performance of an information retrieval model through a set of queries. However, the average set of queries can hide many details. It is not so easy to determine what leads to increas or decreas the average precison. to obtain a better explanation and understanding of the difference between the different models, we carried out a query-by-query analysis.

5.2.2. Results

- Significance of the results obtained
In order to verify the significance of the results obtained, we carried out a test on the query number 62 for which our method majors the methods SimRank and cosine.

.I 62
.T
Fuzzy Requests: An Approach to Weighted Boolean Searches
.A
Bookstein, A.
.W
This article concerns the problem of how to permit a patron to represent the relative importance of various index terms in a Boolean request while retaining the desirable properties of a Boolean system.
The character of classical Boolean systems is reviewed and related to the notion of fuzzy sets. The fuzzy set concept then forms the basis of the concept of a fuzzy request in which weights are assigned to index terms.
Ther properties of such a system are discussed, and it is shown that such systems retain the manipulability of traditional Boolean requests.

.B
(JASIS, Vol. 31, No. 4, July 1980, pp. 240-247)

Fig. 6. Query number 62 of CISI test collection

To evaluate this query, we computed exact precision measures p@5, p@10, p@30, p@100 and p@200 representing respectively, the precision values at the top 5, 10, 30, 100 and 200 documents returned and R-precision.

R-precision: R-precision is defined as precision at cut-off R, where R is the number of relevant documents for the query.

	COS	SimRank	**Our method**
p@5	0	0,08	**0,25**
p@10	0	0,16	**0,5**
p@30	0	0,25	**0,58**
p@100	0,33	0,41	**0,66**
p@200	0,41	0,58	**0,75**
R-Prec	0	0,25	**0,5**

Table 1: improvement in average precision at top n documents returned and R-precision

The following graphic allows to visualize the position of relevant documents retrieved and thus the evolution of this position from one method to another.

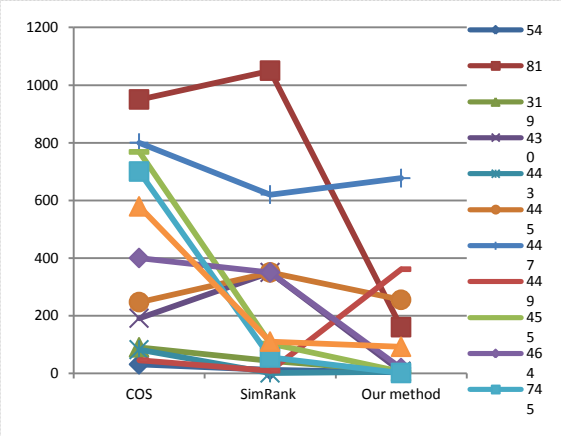

Fig. 7. Comparison of three methods of the ranks of relevant documents to the query number 62

Begin by studying the top relevant documents' of the list:

the documents number 1103 and 54, retrieved in 579th and 31st position by the cosine method, 110th and 12th position by the SimRank method progressing a few ranks through our method (92nd and 3rd).
Similarly the following two relevant documents number 319 and 443 retrieved themselves in 90th and 82nd position and are retrieved improved by SimRank which classifies them into 43rd and the first position and such documents are retrieved topping the list by applying our method (8th and 6th position). This is due to the fact that these documents have a strong direct relationship with the query and an inter significant resemblance increasing their similarity to the query.
The following two relevant documents, 447, and 445, respectively positioned in 800th, 247th by the cosine method, are around 620th and 350th position by the SimRank and respectively in the 675th and 255th by our method, this indicates an indirect resemblance to documents not resembling the query, hence the lightweight distance from the topping list.
The most outstanding gain is obtained by the documents number 430, 81.464, 455 and 745 positioned in191st, 950th, 400th, 768th and 700th by the cosine who find themselves powered the 350th, 1050th, 350th, 103rd and 550th by SimRank and 7th, 160TH, 18th, 5th and 1st by our method.
We suppose this is due to low cosine of these documents with the query and an indirect strong relationship with it.

On 12 relevant documents, 1 document regresses, 2 documents practically keep the same position and 9 documents are progressing in rank relative to SimRank method. However, only one document regresses, 1 document practically keeps the same position and 10 documents are progressing in rank relative to the cosine method.
The documents with a high cosine progress slightly with SimRank, indeed SimRank measure indicates a resemblance in both direct (such as cosine) to which resemblance in both indirect resemblance is added, and it is probable that the documents directly strongly resembling also have strong indirect relationship, which results in our method.

Concerning documents to be very strong progression (430, 81.464, 455et 745, 1103, 443, 319), we are pleasantly surprised to see that indirect resemblances can also significantly bring back the documents to the documents in greater direct resemblance departure.

- *Average Precision of n documents returned*

The evolution of the average Precision of n returned documents (0 <n <200) for Simrank is 0.15 around 0.025 for the Cosine and 0.4108 for our method. So, our method majorises the SimRank and cosine. It therefore appears of interest that we have tried to highlight.
The figure below shows a comparison between the Cosine and Simrank measures and our method in terms of the average precision n per query:

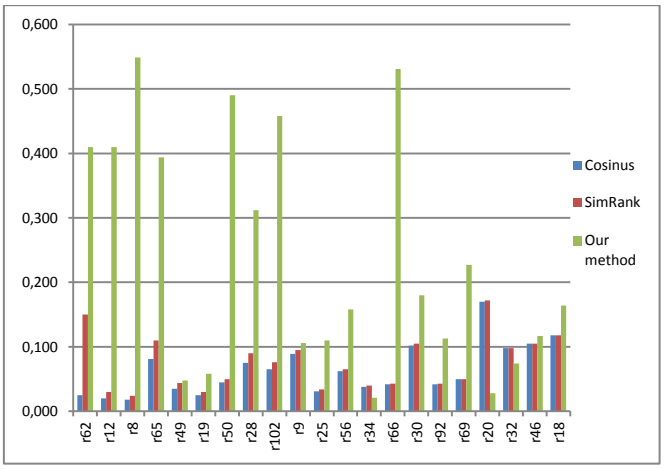

Figure 8. Comparison Cosine / Simrank / our method of the average precision n by request.

For the majority of queries, our method majorises the cosine and SimRank.
This graph shows 22 requests (of 67 treated) for which our method majorises the cosine and SimRank. This confirms that our proposed approach appears to have a positive effect in the majority of cases.

	COS	SimRank	**Our method**
MAP	0,064	0,075	**0,2360952**

Table 2: Mean Average Precision computed over all topics

- The *11-point precision-recall curve*

The 11-point precision-recall curve is a graph plotting the interpolated precision of an information retrieval (IR) system at 11 standard recall levels, that is, {0.0,0.1,0.2,...,1.0}. The method for interpolation is detailed below. The graph is widely used to evaluate IR systems that return ranked documents, which are common in modern search systems.
the *interpolated precision Pinterp* at a certain recall level r is defined as the highest precision found for any recall level $r' \geq r$:

$$p_{interp}(r) = \max_{r' \geq r} p(r')$$

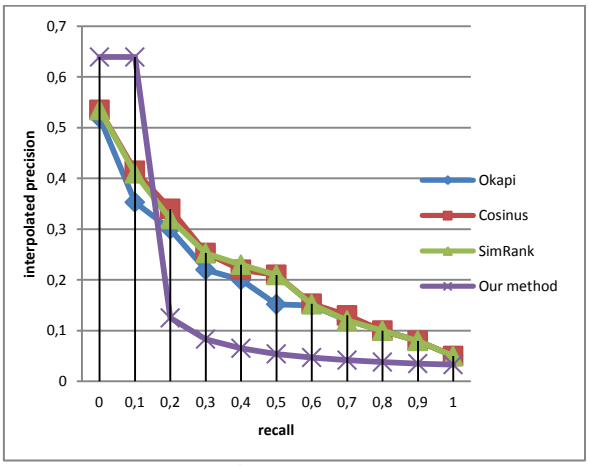

Fig. 9. 11-point precision-recall curve for Cosine, SimRank, Okapi and our method on Cisi

The figure 9 shows that the three measures (Cosine, SimRank, Okapi) obtains similar results. Our method obtains the best scores when the recall is less than 20%. The SimRank achieve the best results when the return rate is between 20% and 30%. The results obtained by the three measures become almost identical when the number of documents returned increases.

Indeed, in our method, when the precision increases, the recall decreases and vice versa. This curve demonstrates that it is be possible to obtain high precision at the cost of low recall or a high recall of a low precision prices. The advantage of this interpolation is that it permits to know the precision to standardized values.

6. Discussion

In this work, we have proposed to cross the language barrier in MIR using a pivot language. This language is used to represent the document and the query regardless of the source and target languages. All the *problem then is the definition of pivot language for MIR* and conversion between natural language and this language.
The LSA algorithm applied to MIR may be seen as the introduction of a pivot language by changing the expression space of index vectors on new dimensions concretizing the pivot : the document and the query are represented by *concept types* in a common space independent of language. This approach is based on the vector model. The *problem posed is to find, regardless of the language, the descriptors or basic unit of information that are identifiable and extractable well as most relevant to document collection written in a given language.*

We described a pivot language based approach to multilingual document representation. We aim to compute concept type and document vectors which creates a semantic space that will serve as a pivot language for information retrieval. This language is used as a semantic indexing base adapted to a trilingual corpus *(*Arabic, French and English). It is an entirely statistics-based, unsupervised, and language independent approach to MIR.
This approach provides an interesting performance for Arabic because words are not explicitly defined in this language.

In our approach, a vector is produced for each document of a corpus, in order to compare them. A document search query is converted into a vector and the vicinity of this query vector in document space constitutes its response.
A document is represented by a vector, where each dimension corresponds to a given concept type and where each value encodes the concept type importance in the document. LSI consist in “projecting” documents on a set of topics learned in an unsupervised manner.
It should be noted that during our learning stage, as a by-product of the projection of the training documents, one also obtains an **embedding of the words** in a typically small dimensional space. The distance between two words in this space translates the measure of similarity between words which is captured by the topic models. For LSI, the implicit measure is the number of co-occurrences in the training corpus.

Our experiments on several document retrieval tasks such as the Quran corpora and the s**econd test collection (CISI)** shows that the method is competitive with state-of-the-art methods. The good performance demonstrates the merits of document Vector in capturing the semantics of documents and descriptors.

We have evaluated the relevance of our approach and of the approach induces in (Yaël, 2009). Our postulate is that our method improves the performances of the multilingual IRS in comparison with structural similarities methods.
The results obtained with the Quran corpora are encouraging and variability shows that they are perfectible. Most queries are processed verses and deal with a specific topic.
In contrast to (McNamee and Mayfield, 2004; Peter et al., 2008), in our approach we proposed to the variable length N-gram extraction from the relevant word groups located between the borders (of the various fragments separated by the border) as this strategy brings an interesting performance for languages (such as Arabic and Chinese) in which the words are not explicitly defined and different words are not separated by spaces, so a sentence is composed of many consecutive characters. each language to its own properties.

On the contrary of (Sami et al., 2009), in our approach we don't use grammatical labels we use an independent language surface analysis.

7. Conclusion and future work

We have presented in this paper an approach for multilingual document representation which combines surface analysis and the LSA statistical algorithm for the detection of concepts in order to create a semantic space that will serve as a pivot language for multilingual information retrieval.
This language is used as a semantic indexing base adapted to trilingual corpus *(*Arabic, French and English) to characterize the documentary content by knowledge , not language-dependent documents .
In our extraction work of n-grams candidates, the words giving less information (i.e. empty words, parasite words and the most frequent words) are replaced by a border.
the words bringing little information ie empty words, noise words and very frequent words are replaced by borders.

In the framework of this work, we were interested in terms extraction methods from parallel corpus. We presented the problem of the construction and definition of a pivot language and we have applied to the Koranic text and the CISI corpus.

The pivot language created can then be used in improving information retrieval, indexing and machine translation or any other applications in the multilingual Web.

References

Emad Elabd, Eissa Alshari, and Hatem Abdulkader. Semantic Boolean Arabic Information Retrieval. The International Arab Journal of Information Technology, Vol. 12, No. 3, May 2015.

Rekha Warrier, Sharvari.S.Govilkar. Cross Language Information Retrieval using Multilingual Ontology as Translation and Query Expansion Base. International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE). Volume 5, Issue 8,August 2015.

Roussey Cathrine. Thèse pour obtenir le grade de docteur. Une méthode d'indexation sémantique adapté aux corpus multilingues, décembre 2001.

Pham Trong Ton, Jean-Pierre Chevallet, Lim Joo Hwee, *Fusion de multi-modalités et réduction par sémantique latente, Laboratoire Image Perception Access & Language* » (IPAL) - UMI CNRS 2955 21 Heng Mui Keng Terrace, 119613, Singapore, 2008.

Susan T. Dumais, Thomas K. Landauer, Michael L. Littman - Automatic Cross-Lingustic Information Retrieval using Latent Semantic Indexing - *Proceedings of SIGIR'96, 1996.*

McNamee and J. Mayfield. Character *N*Gram Tokenization for European Language Text Retrieval. *Information Retrieval* 7, 73-97, 2004

Sami Laroum, Nicolas Béchet, Hatem Hamza et Mathieu Roche, Classification automatique de documents bruités à faible contenu textuel, Manuscrit auteur, publié dans "RNTI : Revue des Nouvelles Technologies de l'Information(2009).

Bourigault, D. (1994). LEXTER un Logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes. Thèse en mathématiques, informatique appliquée aux sciences de l'homme, École des hautes Études en sciences sociales, Paris.

Yaël Champclaux, Un modèle de recherche d'information basé sur les graphes et les similarités structurelles pour l'amélioration du processus de recherche d'information, 2009.

Glen Jeh and Jennifer Widom. *SimRank: A measure of structural-context similarity*. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada, July 2002.

Rémi Lebret, Ronan Collobert. N-gram based low-dimensional representation for document classification. Under review as a conference paper at ICLR 2015.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In NIPS. 2013.

Pennington, J., Socher, R., and Manning, C. D. GloVe: Global Vectors for Word Representation. In Proceedings of EMNLP, 2014.

Paul Bloom. Precis of how children learn the meanings of words. Behavioral and Brain Sciences, 24:1095–1103, 2001.

D. Roy. Grounded spoken language acquisition: Experiments in word learning. IEEE Transactions on Multimedia, 5(2):197–209, June 2003. ISSN 1520-9210. doi: 10.1109/TMM. . 811618. 2003.

Karl Moritz Hermann and Phil Blunsom. Multilingual Distributed Representations without Word Alignment. proceeding of ICLR. arXiv:1312.6173v4 [cs.CL], 20 Mar 2014a.

Karl Moritz Hermann and Phil Blunsom. Multilingual Models for Compositional Distributional Semantics. In ACL. 2014b.

Tomas Kocisky, Karl Moritz Hermann, and Phil Blunsom. Learning Bilingual Word Representations by Marginalizing Alignments. In ACL. 2014.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. Bilbowa: Fast bilingual distributed representations without word alignments. In NIPS Deep Learning Workshop. 2014.

Hieu Pham, Minh-Thang Luong and Christopher D. Manning. Learning Distributed Representations for Multilingual Text Sequences. Proceedings of NAACL-HLT 2015, pages 88–94, Denver, Colorado, May 31 – June 5, 2015. c2015 Association for Computational Linguistics.

Sarath Chandar A P, Stanislas Lauly, Hugo Larochelle, Mitesh Khapra, Balaraman Ravindran, Vikas C Raykar, and Amrita Saha. An autoencoder approach to learning bilingual word representations. In NIPS. 2014.

Minh-Thang Luong, Hieu Pham and Christopher D. Manning. Bilingual Word Representations with Monolingual Quality in Mind. Proceedings of NAACL-HLT 2015, pages 151–159, Denver, Colorado, May 31 – June 5, 2015. c2015 Association for Computational Linguistics.

Stéphane Clinchant and Florent Perronnin. Aggregating ContinuousWord Embeddings for Information Retrieval. Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality, pages 100–109, Sofia, Bulgaria, August 9 2013. c2013 Association for Computational Linguistics.

Aliane Hassina. An ontology based approach to multilingual information retrieval. proceeding of ICTTA'06 SYRIA, 2006.

Peter A. Chew, Brett W. Bader, Ahmed Abdelali, « Latent Morpho-Semantic Analysis: Multilingual Information Retrieval with Character N-Grams and Mutual Information » Proceedings of the 22nd International Conference on Computational Linguistics, August 2008.

Yik-Cheung Tam and Tanja Schultz. 2007. Bilingual LSA-based translation lexicon adaptation for spoken language translation. In Interspeech.

Nick Ruiz and Marcello Federico. 2011. Topic adaptation for lecture translation through bilingual latent semantic models. In WMT.

Bing Zhao and Eric P. Xing. HM-BiTAM: Bilingual Topic Exploration, Word Alignment, and Translation. In NIPS. 2007.

XiaochuanNi, Jian-Tao Sun, Jian Hu, and Zheng Chen. Mining multilingual topics from wikipedia. In WWW. 2009.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. Polylingual topic models. In EMNLP. 2009.

Ivan Vulic,Wim De Smet, Marie-FrancineMoens, and KU Leuven. Identifying word translations from comparable corpora using latent topic models. In ACL-HLT. 2011.
S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman. 1988. Using latent semantic analysis to improve access to textual information. In CHI.

Balpe, J.P., Lelu, A. Papy, F., *Techniques avancées pour l'hypertexte*. Paris, Hermes. 1996.

C. E. Shannon, "Prediction and entropy of printed English," *Bell System Technical Journal* 30, pages 50 - 64. http://languagelog.ldc.upenn.edu/myl/Shannon1950.pdf. 1950.

Radwan Jalam1, Jean-Hugues Chauchat, Pourquoi les n-grammes permettent de classer des textes? Recherche de mots-clefs pertinents à l'aide des n-grammes caractéristiques, 6es Journées internationales d'Analyse statistique des Données Textuelles. JADT 2002.

F. Cuna Ekmekcioglu, Michael F. Lynch, and Peter Willett, "Stemming and N-gram Matching For Term Conflation In Turkish Texts," available at http://www.shef.ac.uk/uni/academic/I-M/is/lecturer/paper13.html#lovi68.1996.