THESAURUS BASED SEMANTIC REPRESENTATION IN LANGUAGE MODELING FOR MEDICAL ARTICLE INDEXING

Jihen Majdoubi, Mohamed Tmar and Faiez Gargouri
Multimedia Information System and Advanced Computing Laboratory
Higher Institute of Information Technologie and Multimedia, Sfax, Tunisia

Keywords: Medical article, Conceptual indexing, Language models, MeSH thesaurus.

Abstract: Language modeling approach plays an important role in many areas of natural language processing including speech recognition, machine translation, and information retrieval. In this paper, we propose a contribution for conceptual indexing of medical articles by using the MeSH (Medical Subject Headings) thesaurus, then we propose a tool for indexing medical articles called SIMA (System of Indexing Medical Articles) which uses a language model to extract the MeSH descriptors representing the document. To assess the relevance of a document to a MeSH descriptor, we estimate the probability that the MeSH descriptor would have been generated by language model of this document.

1 INTRODUCTION

The goal of an Information Retrieval System (IRS) is to retrieve relevant information to a user's query. This goal is quite a difficult task with the rapid and increasing development of the Internet.

Indeed, web information retrieval becomes more and more complex for the user which IRS provide a lot of information, but he often fail, to find the best one in the context of his information need.

The classical IRS are not suitable for managing this growing volume of data and finding relevant documents to a user's information need. The information retrieval techniques commonly used are based on statistical methods and do not take into account the meaning of words contained in the user's query as well as in the documents. Indeed, the current IRS use simple keyword matching: a document to be returned to the user, should contain at least one word of the query. However a document can be relevant even it does not contain any word of the query.

As a simple example, if the query is about "operating system", a document containing windows, unix, vista, and not the term "operating system", would not be retrieved by classical search engines. Consequently, the recall is often low.

Thus, much more "intelligence" should be embedded to IRS in order to be to understand the meaning of the word.

Adding a semantic resource (dictionaries, thesaurus, domain ontologies) to IRS is a possible solution to this problem of the current web.

As in the example of "operating system" cited above, by using concepts of the semantic resource (SR) and their description, the IRS can detect the relationships between operating system, windows, unix, vista and return the document that mentions windows as an answer to the query about "operating system".

Consequently, incorporating the semantic in the IR process can improve the IRS performance.

In the literature, there are three main approaches regarding the incorporation of semantic information into IRS: (1) semantic indexing, (2) conceptual indexing and (3) query expansion.

1. Semantic indexing (Sense Based Indexing): is an indexing approach based on the word senses. The basic idea is to index word meanings, rather than words taken as lexical strings.
   For example, bank (river/money) and plant (manufacturing/life) (Sanderson.M, 1994)(yarowski.D, 1993).
   Thus, word Sense Disambiguation (WSD) algorithms are needed in order to resolve word ambiguity in the document and determine its best word sense.
   The usage of word senses in the process of document indexing is an issue of discussions.
   (Gonzalo.J et al., 1998) performed experiments in sense based indexing: they used the SMART

Majdoubi J., Tmar M. and Gargouri F. (2010).
THESAURUS BASED SEMANTIC REPRESENTATION IN LANGUAGE MODELING FOR MEDICAL ARTICLE INDEXING.
In *Proceedings of the 12th International Conference on Enterprise Information Systems - Artificial Intelligence and Decision Support Systems*, pages 65-74
DOI: 10.5220/0002903300650074

retrieval system and a manually disambiguated collection (Semcor). The results of their experiments proved that indexing by synsets can increase recall up to 29 % compared to word based indexing.
Ellen Voorhees (Voorhees.E.M, 1998) applied word meanings indexing. in the collection of documents, as well as in the query. Comparing the results obtained with the performance of a standard run, (Voorhees.E.M, 1998) affirmed than the overall results have shown a degradation in IR effectiveness when word meanings were used for indexing. She states that a long query has a bad influence on these results and degrades the IR performance.

2. Conceptual Indexing. Unlike previous indexing systems that use lists of simple words to index a document, conceptual indexing is based on concepts issued from the SR.
The conceptual indexing technique has been used in several works (Baziz.M, 2006) (Stairmand.A and J.William, 1996) (Mauldin.M.L, 1991). However, to our knowledge, the most intensive work in this direction was performed by Woods (Woods.W.A, 1997) that proposed an approach which was evaluated using small collections, as for example the unix manual pages (about 10MB of text). To evaluate his system, he defines a new measure, called success rate which indicates if a question has an answer in the top ten documents returned by a retrieval system. The success rate obtained was 60% compared to a maximum of 45% obtained using other retrieval systems.
The experiments described in (Woods.W.A, 1997) are based on small collections of text. But, as shown in (Ambroziak.J, 1997), this is not a limitation; conceptual indexing can be successfully applied to much larger text collections.
3. Query Expansion. SR can also help the user to choose search terms and formulate its requests. For example, (Mihalcea.D and Moldovan, 2000) and (Voorhees.E.M, 1994) propose an IRS which use a thesaurus WordNet to expand the user's query. Such as the query is expanded with terms similar to those of the original query.

These studies showed that IRS based either on conceptual indexing, semantic indexing or a query expansion can improve the effectiveness of IRS.

In our work, we are interested in the conceptual indexing. The essential argument which motivates our choice is that we are concerned about the medical field, and that the technique of conceptual indexing have been used with success in particular domains, suchas the legal field (Stein.J.A, 1997), medical field (Muller.H et al., 2004) and sport field (Khan.L, 2000).

In this paper, we propose our contribution for conceptual indexing of medical articles by using the language modeling approach.

After summarizing the background for this problem in the next section, we present the previous work according to indexing medical articles in section 3. Section 4 explains the language model for Information Retrieval. Following that, we detail our conceptual indexing approach in Section 5. An experimental evaluation and comparison results are discussed in sections 6 and 7. Finally section 8 presents some conclusions and future work directions.

2 BACKGROUND

2.1 Context

Each year, the rate of publication of scientific literature grows, making it increasingly harder for researchers to keep up with novel relevant published work. In recent years big efforts have been devoted to attempt to manage effectively this huge volume of information, in several fields.

In the medical field, scientific articles represent a very important source of knowledge for researchers of this domain. The researcher usually needs to deal with a large amount of scientific and technical articles for checking, validating and enriching of his research work.

This kind of information is often present in electronic biomedical resources available through the Internet like CISMEF[1] and PUBMED[2]. However, the effort that the user put into the search is often forgoten and lost.

To solve these issues, current health Information Systems must take advantage of recent advances in knowledge representation and management areas such as the use of medical terminology resources. Indeed, these resources aim at establishing the representations of knowledge through which the computer can handle the semantic information.

2.2 Medical Terminology Resources

The language of biomedical texts, like all natural language, is complex and poses problems of synonymy and polysemy. Therefore, many terminological systems have been proposed and developed such

[1] http://www.chu-rouen.fr/cismef/

[2] http://www.ncbi.nlm.nih.gov/pubmed/

as Galen, UMLS, GO and MeSH.
In this section, we present some examples of medical terminology resources:

- SNOMED is a coding system, controlled vocabulary, classification system and thesaurus. It is a comprehensive clinical terminology; designed to capture information about a patient's history, illnesses, treatment and outcomes.
- Galen[3] (General Architecture for Language and Nomenclatures) is a system dedicated to the development of ontology in all medical domains including surgical procedures.
- The Gene Ontology is a controlled vocabulary that covers three domains:
  - cellular component, the parts of a cell or its extra cellular environment,
  - molecular function, the elemental activities of a gene product at the molecular level, such as binding or catalysis,
  - biological process, operations or sets of molecular events
- The Unified Medical Language System (UMLS) project was initiated in 1986 by the U.S. National Library of Medicine (NLM). It consists of a (1) metathesaurus which collects millions of terms belonging to nomenclatures and terminologies defined in the biomedical domain and (2) a semantic network which consists of 135 semantic types and 54 relationships.
- The Medical Subject Headings (MeSH) [4] thesaurus is a controlled vocabulary produced by the National Library of Medicine (NLM) and used for indexing, and searching for biomedical and health-related information and documents.

Us for us, we have chosen Mesh because it meets the aims of medical librarians and it is a successful tool and widely used for indexing literature.

3 PREVIOUS WORK

Automatic indexing of the medical articles has been investigated by several researchers. In this section, we are only interested in the indexing approach using the MeSH thesaurus.

(N*é*v*é*ol.A, 2005) proposes a tool called MAIF (MesH Automatic Indexer for French) which is developed within the CISMeF team. To index a medical ressource, MAIF follows three steps: analysis of the resource to be indexed, translation of the emerging concepts into the appropriate controlled vocabulary (MeSH thesaurus) and revision of the resulting index.

In (Aronson.A et al., 2004), the authors proposed the MTI (MeSH Terminology Indexer) used by NLM to index English resources. MTI results from the combination of two MeSH Indexing methods: MetaMap Indexing (MMI) and a statistical, knowledge-based approach called PubMed Related Citations (PRC).

The MMI method (Aronson.A, 2001) consists of discovering the Unified Medical Language System(UMLS) concepts from the text. These UMLS concepts are then refined into MeSH terms.

The PRC method (Kim.W et al., 2001) computes a ranked list of MeSH terms for a given title and abstract by finding the MEDLINE citations most closely related to the text based on the words shared by both representations.

Then, MTI combines the results of both methods by performing a specific post processing task, to obtain a first list. This list is then devoted to a set of rules designed to filter out irrelevant concepts. To do so, MTI provides three levels of filtering depending on precision and recall: the strict filtering, the medium filtering and the base filtering.

Nomindex (Pouliquen.B, 2002) recognizes concepts in a sentence and uses them to create a database allowing to retrirve documents. Nomindex uses a lexicon derived from the ADM (Assisted Medical Diagnosis) (Lenoir.P et al., 1981) which contains 130.000 terms.

First, document words are mapped to ADM terms and reduced to reference words. Then, ADM terms are mapped to the equivalent French MeSH terms, and also to their UMLS Concept Unique Identifier. Each reference word of the document is then associated with its corresponding UMLS. Finally a relevance score is computed for each concept extracted from the document.

(N*é*v*é*ol.A et al., 2007) showed that the indexing tools cited above by using the controlled vocabulary MeSH, increase retrieval performance.

These approaches are based on the vector space model. We propose in this paper our tool for the medical article indexing which is based on the language modeling.

[3] http://www.opengalen.org

[4] http://www.nlm.nih.gov/mesh/

4 THE LANGUAGE MODELING BASED INFORMATION RETRIEVAL

Language modeling approachs to information retrieval are attractive and promising because they rely to the problem of information retrieval with that of language model estimation, which has been studied extensively in other application areas such as recognition.

Many approaches of language modeling has been used in information retrieval (Ponte.M and Croft, 1998)(Lafferty.J and Zhai, 2001).

The basic idea of these approaches is to estimate a language model for each document D in the collection C, and then to rank documents by the likelihood of the query according to the estimated language model.

Each query Q is treated as a sequence of independent terms ($Q = \{q_1, q_2, \ldots, q_n\}$). Thus, the probability of generating Q having document D can be obtained and retrieved documents are ranked according to it:

$$P(Q|D) = \prod_{q_i \in Q} P(q_i|D)$$

where q_i is the i^{th} term in the query.

It is important to note that non-zero probability should be assigned to query terms that do not appear in a given document. Thus language models for information retrieval must be ”smoothed”.

There are many smoothing approachs for language modeling in IR. A popular approach combines a component estimated from the document and another from the collection by linear interpolation:

$$P(t|D) = \lambda P_{doc}(t|D) + (1-\lambda) P_{coll}(t)$$

where $\lambda \in [0,1]$ is a weighting parameter.

Language modeling approach show a signifiant effectivness to information retrieval. However most parameter estimation approaches in language model do not consider the semantic: the probability of term t is only the combination of distributions in the document and the corpus of that word itself. In fact, the document that contains the term ”car” can not be retrieved to answer a query containing ”automobile”, if this query term is not present in the document.

Thus in order to bring semantic feature into language model, semantic smoothing technique is necessary.

Recently, many attempts have been made to enrich language models with more complex syntactic and semantic models, with varying success.

For example, (Lafferty.J and Zhai, 2001) proposed a method capturing semantic relations between words based on term co-occurrences. They use methods from statistical machine translation to incorporate synonymy into the document language model.

(Jin.R et al., 2002) views a title as a translation form that document and the title language model is regarded as an approximate language model of the query and estimated probability under such assumption.

(Zhang.J et al., 2004) proposed a trigger language model based IR system. They compute the associate ratio of the words from training corpus the get the triggered words collection of the query words to find the real meaning of the word in specific text context, which seems to be a variation of computing cooccurrences.

In the next section, we describe our indexing approach based on the semantic language modeling.

5 OUR APPROACH

Our work aims to determine for each document, the most representative MeSH descriptors. For this reason, we have adapted the language model by substituting the query by the Mesh descriptor. Thus, we infer a language model for each document and rank Mesh descriptor according to our probability of producing each one given that model. We would like to estimate $P(Des|M_d)$, the probability of the Mesh descriptor given the language model of document d.

Our indexing methodology as shown by figure 1, consists of three main steps: (a) Pretreatment, (b) concept extraction and (c) generation of the semantic core of document.

We present the architecture components in the following subsections.

5.1 MeSH Thesaurus

The structure of MeSH is centered on descriptors, concepts, and terms.

- Each term can be either a simple or a composed term.
- A concept is viewed as a class of synonymous terms, one of then (called Preferred term) gives its name to the concept.
- A descriptor class consists of one or more concepts where each one is closely related to each other in meaning.
  Each descriptor has a preferred concept. The descriptor’s name is the name of the preferred concept. Each of the subordinate concepts is re-

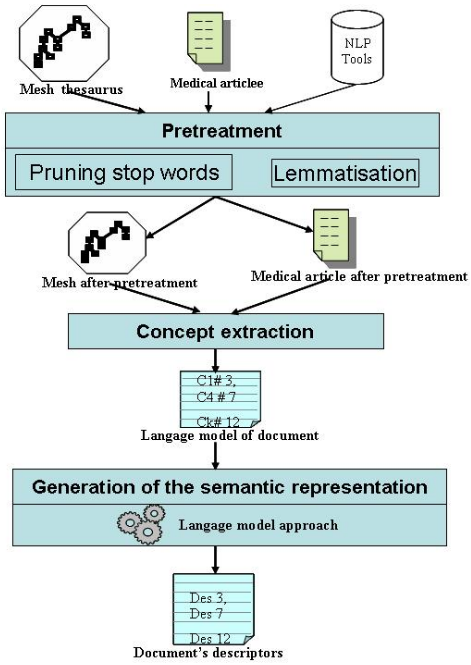
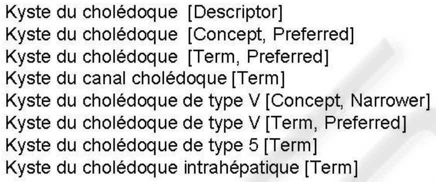


Figure 1: Architecture of our proposed approach.

lated to the preferred concept by a relationship (broader, narrower).

Its important to note that the Descriptors MeSH are also interconnected by the relationship "related".



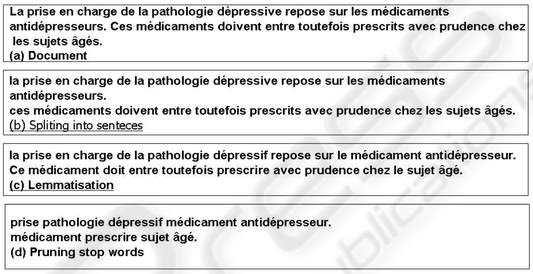
Kyste du cholédoque [Descriptor]
Kyste du cholédoque [Concept, Preferred]
Kyste du cholédoque [Term, Preferred]
Kyste du canal cholédoque [Term]
Kyste du cholédoque de type V [Concept, Narrower]
Kyste du cholédoque de type V [Term, Preferred]
Kyste du cholédoque de type 5 [Term]
Kyste du cholédoque intrahépatique [Term]

Figure 2: Extrait of MeSH.

As shown by figure 2, the descriptor "*Kyste du cholédoque*" consists of two concepts and five terms. The descriptor's name is the name of its preferred concept. Each concept has a preferred term, which is also said to be the name of the Concept. For example, the concept "*Kyste du cholédoque*" has two terms "*Kyste du cholédoque*" (preferred term) and "*Kyste du canal cholédoque*". As in the example above, the concept "*Kyste du choldoque de type V*" is narrower to than the preferred concept "*Kyste du canal cholédoque*".

5.2 Pretreatment

The first step is to split text into a set of sentences. We use the Tokeniser module of GATE (Cunningham.M et al., 2002) in order to split the document into tokens, such as numbers, punctuation, character and words. Then, the TreeTagger (Schmid.H, 1994a) stems these tokens to assign a grammatical category (noun, verb...) and lemma to each token. Finally, our system prunes the stop words for each medical article of the corpus. This process is also carried out on the MeSH thesaurus. Thus, the output of this stage consists of two sets. The first set is the article's lemma, and the second one is the list of lemma existing in the MeSH thesaurus. Figure 3 outlines the basic steps of the pretreatment phase.



La prise en charge de la pathologie dépressive repose sur les médicaments antidépresseurs. Ces médicaments doivent entre toutefois prescrits avec prudence chez les sujets âgés.
(a) Document

la prise en charge de la pathologie dépressive repose sur les médicaments antidépresseurs.
ces médicaments doivent entre toutefois prescrits avec prudence chez les sujets âgés.
(b) Spliting into senteces

la prise en charge de la pathologie dépressif repose sur le médicament antidépresseur.
Ce médicament doit entre toutefois prescrire avec prudence chez le sujet âgé.
(c) Lemmatisation

prise pathologie dépressif médicament antidépresseur.
médicament prescrire sujet âgé.
(d) Pruning stop words

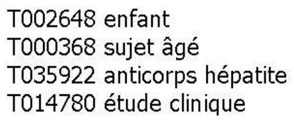
Figure 3: Pretreatment step.

5.3 Concept Extraction

This step consists of extracting single word and multiword terms from texts that correspond to MeSH concepts. So, SIMA processes the medical article sentence by sentence. Indeed, in the pretreatment step, each lemmatized sentence S is represented by a list of lemma ordered in S as they appear in the medical article. Also, each MeSH term t_j is processed with TreeTagger in order to return its canonical form or lemma. Let: $S = (l_1, l_2, \ldots, l_n)$ and $\big(t_j = (att_{j1}, att_{j2}, \ldots, att_{jk})\big)$. The terms of a sentence S_i are:

$$Terms(S_i) = \big\{T, \forall att \in T, \exists l_{ij} \in S_i, att = l_{ij}\big\}$$

For example, let us consider the lemmatized sentence S_1 given by:
$S_1 = (étude, enfant, agé, sujet, anticorps, virus, hépatite\}$.



T002648 enfant
T000368 sujet âgé
T035922 anticorps hépatite
T014780 étude clinique

Figure 4: Example of terms.

If we consider the set of terms shown by figure 4, this sentence contains three different terms: (i) *enfant*, (ii) *sujet agé* and (iii) *anticorps hépatite*. The term *étude clinique* is not identified because the word *clinique* is not present in the sentence S_1.

Thus:

$Terms(S_1) = \{enfant, sujet\ agé, anticorps\ hépatite\}$.

A concept c_i is proposed to the system like a concept associated to the sentence S ($Concepts(S)$), if at least one of its terms belongs to $Terms(S)$.

For a document d composed of n sentences, we define its concepts (Concepts(d)) as follows:

$$Concepts(d) = \bigcup_{i=1}^{n} Concepts(S_i) \qquad (1)$$

Given a concept c_i of $Concepts(d)$, its frequency in a document d ($f(c_i,d)$) is equal to the number of sentences where the concept is designated as $Concepts(S)$. Formally:

$$f(c_i,d) = \left| \sum_{c_i \in Concepts(S_j)} S_j \in d \right| \qquad (2)$$

5.4 Generation of the Semantic Core of Document

To determine the MeSH descriptors from documents, we estimated a language model for each document in the collection and for a MeSH descriptor we rank the documents with respect to the likelihood that the document language model generates the MeSH descriptor. This can be viewed as estimating the probability $P(d|des)$.

To do so, we used the language model approach proposed by (Hiemstra.D, 2001).

For a collection D, document d and MeSH descriptor (des) composed of n concepts:

$$P(d|des) = P(d). \prod_{c_j \in des} (1-\lambda).P(c_j|D) + \lambda.P(c_j|d) \qquad (3)$$

We need to estimate three probabilities:

1. $P(d)$: the prior probability of the document d:

$$P(d) = \frac{|concepts(d)|}{\sum_{d' \in D} |concepts(d')|} \qquad (4)$$

2. $P(c|D)$: the probability of observing the concept c in the collection D:

$$P(c|D) = \frac{f(c,D)}{\sum_{c' \in D} f(c',D)} \qquad (5)$$

where $f(c,D)$ is the frequency of the concept c in the collection D.

3. $P(c|d)$: the probability of observing a concept c in a document d:

$$P(c|d) = \frac{cf(c,d)}{|concepts(d)|} \qquad (6)$$

Several methods for concept frequency computation have been proposed in the literature. In our approach, we applied the weighting concepts method (CF: Concept Frequency) proposed by (Baziz.M, 2006).

So, for a concept c composed of n words, its frequency in a document depends on the frequency of the concept itself, and the frequency of each sub-concept. Formally:

$$cf(c,d) = f(c,d) + \sum_{sc \in subconcepts(c)} \frac{length(sc)}{length(c)}.f(sc,d) \qquad (7)$$

with:

- $Length(c)$ represents the number of words in the concept c.
- $subconcepts(c)$ is the set of all possible concepts MeSH which can be derived from c.

For example, if we consider a concept ”bacillus anthracis”, knowing that ”bacillus” is itself also a MeSH concept, its frequency is computed as:
$cf(bacillus\ anthracis) = f(bacillus\ anthracis) + \frac{1}{2}.f(bacillus)$
consequently:

$$P(d|des) = \frac{|concepts(d)|}{\sum_{concepts(d') \in D} |concepts(d')|} . \prod_{c \in des} \left((1-\lambda).\frac{f(c,D)}{\sum_{c' \in D} f(c',D)} + \lambda. \left(\frac{f(c,d) + \sum_{sc \in subconcepts(c)} \frac{length(sc)}{length(c)}.f(sc,d)}{|concepts(d)|} \right) \right. \qquad (8)$$

To calcultate the $f(c,d)$, we used the measure CSW that we have defined in (Majdoubi.J et al., 2009).

The measure $CSW(ContentStructureWeight)$ takes into account the concept frequency and the location of each one of its occurrences.

For example, the concept of the ”Title” receives a high importance ($*10$) compared to the concept of the ”Abstract” ($*8$) or of the ”Paragraphs” ($*2$). The various coefficients used to weigh the concept locations were determined in an experimental way in (Gamet.J, 1998). Formally:

$$f(c_i,d) = CSW(c_i,d) = \sum_{c \in A} f(c_i,d,A) \times W_A \qquad (9)$$

Where:

- $f(c_i,D,A)$: the occurrence frequency of the concept c_i in document d at location A,
- $A \in \{title, keywords, abstract, paragraphs\}$,
- W_A: weight of the position A.

consequently:

$$P(c|d) = \frac{CSW(c,d) + \sum_{sc \in subconcepts(c)} \frac{length(sc)}{length(c)} .CSW(sc,d)}{|concepts(d)|}$$

$$= \frac{\sum_{c \in A} f(c,d,A) \times W_A}{|concepts(d)|} +$$

$$\frac{\sum_{sc \in subconcepts(c)} \frac{length(sc)}{length(c)} \cdot \sum_{sc \in A} f(sc,d,A) \times W_A}{|concepts(d)|} \tag{10}$$

$$P(d|des) = \frac{|concepts(d)|}{\sum_{d' \in D} |concepts(d')|} \cdot \prod_{c_j \in des} [(1-\lambda) \cdot \frac{f(c_j,D)}{\sum_{c' \in D} f(c',D)}$$

$$+\lambda. \left(\frac{CSW(c_j,d) + \sum_{sc \in subconcepts(c_j)} \frac{length(sc)}{length(c)} .CSW(sc,d)}{|concepts(d)|} \right)] \tag{11}$$

It is important to note that each descriptor is treated independently of others descriptors: any consideration of the semantic in the calculation of $P(d|des)$ is taken into account. However, as mentionned above the MeSH descriptors are interconnected by the relationship ”related”.
This observation shows that it is necessary to incorporate a kind of semantic smoothing into the calculation of $P(d|des)$.
To do so, we use the function $DescRelatedto_E$ that associates for a given descriptor des_i, all MeSH descriptors relating to des_i among the set of descriptors E.

Thus:

$$P(d_k|des_i) = P(d_k). \prod_{c_j \in des_i} ((1-\lambda) .P(c_j|D) + \lambda.P(c_j|d_k))$$

$$+ \frac{\sum_{g \in DescRelatedto_{descriptorsof(d_k)}(des_i)} P(d_k|g)}{|DescRelatedto_{descriptorsof(d_k)}(des_i)|} \tag{12}$$

Where $descriptorsof(d_k)$ presents the set of MeSH descriptors having a positive probability $P(des|d_k)$ with the document d_k.

$$descriptorsof(d_k) = \{DES, P(des|d_k) > 0\}$$

Finally:

$$P(d_k|des) \quad = \quad \frac{|concepts(d_k)|}{\sum_{d' \in D} |concepts(d')|} \quad \cdot \prod_{c_j \in des} \left[(1-\lambda) \cdot \frac{f(c_j,D)}{\sum_{c' \in D} f(c',D)} \right.$$

$$+\lambda. \left(\frac{CSW(c_j,d_k) + \sum_{sc \in subconcepts(c_j)} \frac{length(sc)}{length(c)} .CSW(sc,d_k)}{|concepts(d_k)|} \right)]$$

$$+ \frac{\sum_{g \in DescRelatedto_{descriptorsof(d_k)}(des_i)} P(d_k|g)}{|DescRelatedto_E(des_i)|} \tag{13}$$

6 EMPIRICAL RESULTS

To evaluate our indexing approach we built a corpus from 500 randomly selected scientific Articles from CISMEF. Analysis of this corpus revealed about $716,000$ words.
These articles have been manually indexed by somee professional indexers of CISMeF team.

Experimental Process

Our experimental process can be mainly divided in these steps:

- Our process begins by dividing each article into a set of sentences. Then, a lemmatisation of the corpus and the Mesh terms is ensured by TreeTagger(Schmid.H, 1994b). After that, a filtering step is performed to eliminate the stop words.
- For each sentence S_i, of a test corpus, we determine the set $concepts(S_i)$.
- For a document d and for each MeSH descriptor des_i, we calculate $P(d|des_i)$.
- In the document d, the MeSH descriptors are rankeded by decreasing scores $P(d|des_i)$.

Performance evaluation was done over the same set of 500 articles, by comparing the set of MeSH descriptors retrieved by our system against the manual indexing (presented by the professional indexers).

For this evaluation, Three measures have been used: *precision*, *recall* and *F-measure*.
Precision corresponds to the number of indexing descriptors correctly retrieved over the total number of retrieved descriptors.
Recall corresponds to the number of indexing descriptors correctly retrieved over the total number of escriptors expected.
F-measure combines both precision and recall with an equal weight.

$$Recall = \frac{TP}{TP+FN} \tag{14}$$

$$Precision = \frac{TP}{TP+FP} \tag{15}$$

Where:

- TP: (true positive) is the number of MeSH descriptors were correctly identified by the system and were found in the manual indexing.
- FN: (false negative) is the MeSH descriptors that the system failed to retrieve in the corpus.
- FP: (false positive) is the number of MeSH descriptors that were retrieved by the system but were not found in the manual indexing.

$$F-measure = \frac{1}{\alpha \times \frac{1}{Precision} + (1-\alpha) \times \frac{1}{Recall}} \quad (16)$$

with $\alpha = 0,5$.

Results and Discuss

In the evaluation process, 3 cases are experimented:

1. case 1: classical langage model: frequency of the concept is calculated by using the equation number 8.
2. case 2: classical langage model+CSW measure: frequency of the concept is calculated by using the CSW measure (see equation 11).
3. case 3: semantic langage model+CSW: using the semantic smoothing combined to CSW measure (see equation 13).

Table 1 shows the precision (P) and the recall (R) obtained by our system SIMA at fixed ranks 1 through 10 in each case cited above.

Table 1: Precision and recall of SIMA at fixed ranks.

Rank	case 1(P/R)	case 2(P/R)	case 3(P/R)
1	46,78/27,42	62,37/32,26	78,11/41,03
4	30,32/33,65	57,21/42,52	67,88/52,31
10	21,23/42,76	48,56/57,39	58,39/74,25

Figure 5 presents the obtained F-measure by our system "SIMA", in the three cases: (i) classical langage model, (ii) classical langage model+CSW measure and (iii) semantic langage model+CSW.

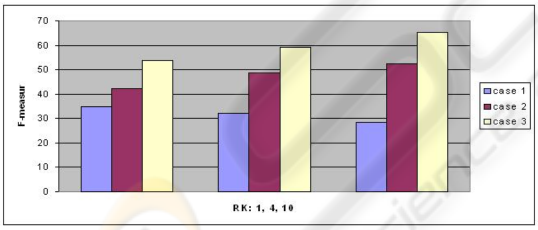


Figure 5: Global comparison results on the three cases.

Results presented in figure 5 clearly show the advantage of using semantic langage model combined to CSW measure for enhancing system performance.

For example, the F-measure value in the rank 4 is $31,89$ in the case of classical langage model, $48,78$ in the case of classical langage model+CSW measure and $59,08$ in the case of semantic langage model combined to CSW measure.

We can also remark that the precision and recall are grower in the case of "semantic langage model+CSW" at all the precision and recall points.

For example, the precision in the rank 4 is $30,32$ in the case of "classical langage model" and $57,21$ (+40% in the case of "classical langage model+CSW measure". It grows to $67,88$ when "semantic langage model+CSW".

The obtained results confirm the well interest to use the third case "semantic langage model+CSW". Taking into account these results, we choose the third case as the best and we have adopted it in the remaining experimentations.

7 COMPARISON OF SIMA WITH OTHERS TOOLS

Encouraged by the previous validation results, we then carry out an experiment which compares SIMA with two MeSH indexing systems: MAIF (MeSH Automatic Indexing for French) and NOMINDEX presented in the section 3.

For this evaluation, we used the same corpus used by (N*é*v*é*ol.A et al., 2005) composed of 82 resources randomly selected in the CISMeF catalogue. It contains about 235,000 words altogether, which represents about 1.7 Mb.

Table 2 shows the precision and recall, obtained by NOMINDEX, MAIF and SIMA at ranks 1,4, 10 and 50 on the test Corpus.

Table 2: Precision and recall, obtained by NOMINDEX, MAIF and SIMA.

Rank	NOMINDEX	MAIF (P/R)	SIMA (P/R)
1	13,25/2,37	45,78/7,42	39,76/6,93
4	12,65/9,20	30,72/22,05	28,53/27,02
10	12,53/22,55	21,23/37,26	20,48/39,42
50	6,20/51,44	7,04/48,50	9,25/40,01

The comparison chart is shown in Figure 6.

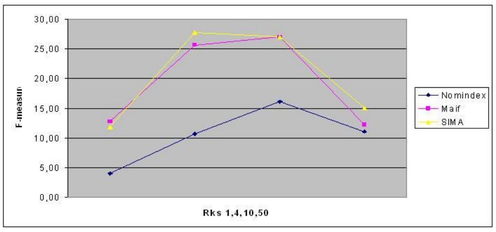


Figure 6: F-measure generated by SIMA compared to NOMINDEX and MAIF.

By examining the figure 6, we can notice that the least effective results come from NOMINDEX with a value of F-measure equal to $4,02$ in rank 1, $10,65$ in rank 4, $16,11$ in rank 10 and $11,07$ in rank 50.

As we can see in figure 6, SIMA and MAIF echoed very similar performance in ranks 1 and 10 with a slight performance. For example, at rank 10

MAIF give the best results with a value of F-measure equal to $27,04$. Concerning SIMA, it generates $26,95$ as value of F-measure at rank 10.

At rank 4, SIMA displayed the best performance results with a F-measure rate of $27,75\%$.

Concerning rank 50, the best result was scored by SIMA with $9,25$ for precision and $15,02$ for F-measure. Regarding MAIF, even though the precision obtained $(7,04)$ is the highest one, its F-measure have been less than SIMA.

8 CONCLUSIONS

The work developed in this paper outlined a concept language model using the Mesh thesaurus for representing the semantic content of medical articles.

Our proposed conceptual indexing approach consists of three main steps. At the first step (Pretreatment), being given an article, MeSH thesaurus and the NLP tools, the system extracts two sets: the first is the article's lemma, and the second is the list of lemma existing in the the MeSH thesaurus. At step 2, these sets are used in order to extract the Mesh concepts existing in the document. After that, our system interpret the relevance of a document d to a MeSH descriptor des by measuring the probability of this descriptor to be generated by a document language $(P(d|des_i))$. Finally, the MeSH descriptors are rankeded by decreasing score $P(d|des_i)$.

We can thus summarize our major contribution by: We evaluated the methods using three measures: precision, recall and F-measure. Our experimental evaluation shows the effectiveness of our approach.

REFERENCES

Ambroziak J. (1997). Conceptually assisted web browsing. In *Sixth International World Wide Web conference*, Santa Clara.

Aronson A. (2001). Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *AMIA*, pages 17–21.

Aronson A., J. Mork, C. Gay, S. Humphrey and W. Rogers (2004). The nlm indexing initiative's medical text indexer. In *Medinfo*.

Baziz M. (2006). *Indexation conceptuelle guidée par ontologie pour la recherche d'information*. PhD thesis, Univ. of Paul sabatier.

Cunningham M., D. Maynard, K. Bontcheva and V. Tablan (2002). Gate: A framework and graphical development environment for robust nlp tools and applications. *ACL*.

Gamet J. (1998). *Indexation de pages web*. Report of dea, universit de Nantes.

Gonzalo J., F. Verdejo, I. Chugur and J. Cigarran (1998). Indexing with wordnet synsets can improve text retrieval. In *COLING-ACL '98 Workshop on Usage of Word.Net in Natural Language Processing Systems*, Montreal, Canada.

Hiemstra D. (2001). *Using Language Models for Information Retrieval*. PhD thesis, University of Twente.

Jin R., A. G. Hauptman and C. Zhai (2002). Title language model for information retrieval. In *SIGIR02*, pages 42–48.

Khan L. (2000). *Ontology-based Information Selection*. PhD thesis, Faculty of the Graduate School, University of Southern California.

Kim W., A. Aronson and W. Wilbur (2001). Automatic mesh term assignment and quality assessment. In *AMIA*.

Lafferty J. and Zhai C. (2001). Document language models, query models, and risk minimization for information retrieval. In *SIGIR'01*, pages 111–119.

Lenoir P., R. Michel, C. Frangeul and G. Chales (1981). R*é*alisation, d*é*veloppement et maintenance de la base de donn*é*es a.d.m. In *Médecine informatique*.

Majdoubi J, M. Tmar and F. Gargouri (2009). Using the mesh thesaurus to index a medical article:combination of content, structure and semantics. In *Knowledge-Based and Intelligent Information and Engineering Systems, 13th International Conference, KES'2009*, page 278285.

Mauldin M. L. (1991). Retrieval performance in ferret: a conceptual information retrieval system. In *lSth International A CM-SIGIR Conference on Research and Development in Information Retrieval*, pages 347–355, Chicago.

Mihalcea D. and Moldovan I. (2000). An iterative approach to word sense disambiguation. In *FLAIRS-2000*, pages 219–223, Orlando,.

Muller H., E. Kenny and P. Sternberg (2004). Textpresso: An ontology-based information retrieval and extraction system for biological literature. In *PLoS Biol*.

N*é*v*é*ol A. (2005). *Automatisation des taches documentaires dans un catalogue de santé en ligne*. PhD thesis, Institut National des Sciences Appliques de Rouen.

N*é*v*é*ol A., Mary V., A. Gaudinat, C. Boyer, Rogozan A. and S. Darmoni (2005). A benchmark evaluation of the french mesh indexers. In *10th Conference on Artificial Intelligence in Medicine, AIME 2005*.

N*é*v*é*ol A., S. Pereira, G. Kerdelhu, B. Dahamna, M. Joubert, and S. Darmoni (2007). Evaluation of a simple method for the automatic assignment of mesh descriptors to health resources in a french online catalogue. In *MedInfo*.

Ponte M. and Croft W. (1998). A language modeling approach to information retrieval. In *ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281.

Pouliquen B. (2002). *Indexation de textes médicaux par indexation de concepts, et ses utilisations*. PhD thesis, Universit Rennes 1.

Sanderson M. (1994). Word sense disambiguation and information retrieval. In *17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 142–151.

Schmid H. (1994a). Probabilistic part-of-speech tagging using decision trees. *International Conference on New Methods in Language Processing. Manchester*.

Schmid H. (1994b). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester.

Stairmand A. and William J. (1996). Conceptual and contextual indexing of documents using wordnet-derived lexical chains. In *18th BCS-IRSG Annual Colloquium on Information Retrieval Research*.

Stein J. A. (1997). Alternative methods of indexing legal material: Development of a conceptual index. In *Law Via the Internet 97*, Sydney, Australia.

Voorhees E. M. (1994). Query expansion using lexical-semantic relations. In *17th Annual International ACM SIGIR, Conference on Research and Development in Information Retrieval*, pages 61–69, Dublin, Ireland.

Voorhees E. M. (1998). Using wordnet for text retrieval. In *WordNet, An Electronic Lexical Database*, pages 285–303.

Woods W. A. (1997). Conceptual indexing: A better way to organize knowledge. Technical Report TR-97-61, Digital Equipment Corporation, Sun Mierosysterns Laboratories.

Yarowski D. (1993). One sense per collocation. In *the ARPA Human Language Technology Workshop*.

Zhang J., Min.Q, Sun.L and Sun.Y (2004). An improved language model-based chinese ir system. In *Journal of Chinese Information Processing*, pages 23–29.