

# Biomedical indexing and retrieval system based on language modeling approach

Jihen MAJDOUBI, Hatem LOUKI, Mohamed TMAR and Faiez GARGOURI

Multimedia InfoRmation system and Advanced Computing Laboratory, Higher Institute of Information Technologie and Multimedia, University of sfax, Tunisia

majdoubi\_jihene@yahoo.fr, hatem.loukil@isimsf.rnu.tn,  
mohamed.tmar@isimsf.rnu.tn, faiez.gargouri@fsegs.rnu.tn

## **ABSTRACT**

*In the medical field, scientific articles represent a very important source of knowledge for researchers of this domain. But due to the large volume of scientific articles published on the web, an efficient detection and use of this knowledge is quite a difficult task.*

*In this paper, we propose our contribution for conceptual indexing of medical articles by using the MeSH (Medical Subject Headings) thesaurus. With this in mind, we propose a tool for indexing medical articles called BIOINSY (BIOmedical Indexing SYstem) which uses a language model for selecting the best representative descriptors for each document.*

*The proposed indexing approach was evaluated by intensive experiments. These experiments were conducted on document test collections of real world clinical extracted from scientific collections, namely PUBMED and CLEF. The results generated by these experiments demonstrate the effectiveness of our indexing approach.*

## **KEYWORDS**

*Medical article, conceptual indexing, Language models, MeSH thesaurus, Biomedical Information Retrieval*

## **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 user who IRS provides a lot of information. However the user often fails, to find the best information in the context of his need. Current IRS rely on simple matching keywords from queries to those from documents under the basic assumption of terms independence. 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, but not the term "operating system", would not be retrieved by classical search engines. Consequently, most IRS provide poor search results and the recall is often low.

A suitable solution to this problem is the conceptual indexing process that uses the ontology's concepts as means of normalizing the document vocabulary. Thus, by using concepts of the semantic resource (SR) and their description, IRS become able to understand the meaning of the word and capture the document semantic content. As in the example of "operating system" cited above, through concept recognition, the IRS can detect the relationships between the terms of user's query. Consequently IRS return the document that mentions Windows as an answer to the query about "operating system".

We focus here on the use of the SR to derive the semantic kernels of biomedical documents. In this paper, we propose a novel method for conceptual indexing of medical articles by using the MeSH (Medical Subject Headings) resource. The goal of this work is 2-fold. First, our goal is to build a system which can annotate a medical article with relevant MeSH descriptors. Second, our goal is to use this automatic annotation method to improve the biomedical document retrieval. The remainder of this paper is organized as follows. After summarizing the background for this problem in the next section, we present the previous work according to indexing medical articles in section 3. We detail our conceptual indexing approach in section 4. An experimental evaluation and comparison results are discussed in sections 5 and 6. After that, we try to use our conceptual indexing approach to improve document retrieval by using ImageCLEF med 2007 test collections. 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 several researches have been devoted to attempt to manage effectively this huge volume of information, in many 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<sup>1</sup> and PUBMED<sup>2</sup>. However, the effort that the user put into the search is often forgotten 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 Examples of medical terminology resources**

The language of biomedical texts, like all natural language, is complex and poses problems of synonymy and polysemy. Therefore, several terminological resources are available to cope with the lexical ambiguity and synonymy present in biomedical terminology.

---

<sup>1</sup> <http://www.chu-rouen.fr/cismef/>

<sup>2</sup> <http://www.ncbi.nlm.nih.gov/pubmed/>

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 (General Architecture for Language and Nomenclatures)<sup>3</sup> 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) thesaurus<sup>4</sup> 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. RELATED RESEARCH 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.

In [2], *Névél* proposes a tool called MAIF (MesH Automatic Indexer for French) which is developed within the CISMéF team. To index a medical resource, 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 [4], the authors proposed the MTI (MeSH Terminology Indexer) 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 [5] consists on discovering the Unified Medical Language System (UMLS) concepts from the text. These UMLS concepts are then refined into MeSH terms.

The PRC method [6] 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

---

<sup>3</sup> <http://www.opengalen.org>

<sup>4</sup> <http://www.nlm.nih.gov/mesh/>

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 [1] recognizes concepts in a sentence and uses them to create a database allowing to retrieve documents. Nomindex uses a lexicon derived from the ADM (Assisted Medical Diagnosis) [7] 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.

The conceptual indexing strategy proposed by [3] involves three steps. First they compute for each concept MeSH  $C$  its similarity with the document  $D$ . After that, the candidate concepts extracted from step 1 are re-ranked according to a correlation measure that estimates how much the word order of a MeSH entry is correlated to the order of words in the document. Finally the content based similarity and the correlation between the concept  $C$  and the document  $D$  are combined in order to compute the overall relevance score. The  $N$  top ranked concepts having the highest scores are selected as candidate concepts of the document  $D$ .

[32] introduced a retrieval-based system for MeSH Classification called EAGL. For each MeSH term, its synonyms and description are indexed as a single document in a retrieval index. A piece of text, the query to the retrieval system, is classified with the best ranked MeSH 'documents'.

KNN (K-Nearest-Neighbours) indexing system [33] relies on a retrieval approach based on language models. The parameters of the query language model are estimated on the text to index. Next, citations most similar to this query language model are retrieved. The classification is based on the MeSH terms assigned to the top  $K^5$  retrieved documents. The relevance of a MeSH term is determined by summing the retrieval score of the top documents that have been assigned that term.

Our indexing approach presented in this paper differs from previous works in the following key points:

– Weighing term: to our knowledge, there is so far no work dealing with the use of the semantics relationships in the weighing process for biomedical IR. Indeed, to determine a term importance, the indexing approaches are based on the statistical measure (Tf, idf). So, any consideration of the term semantic in the calculation of its weight is taken into account. For example, the term weight is calculated independently of its synonym occurrences. However, [8] has showed the utility of integrating the WordNet's conceptual information to calculate term importance for web information retrieval. Thus, motivated by the Search results of [8], we estimate term relevance for a document by using its semantics relationships. To determine term importance, our indexing strategy exploits its relationships in Mesh, rather than relying only on a statistical measure.

– Word Sense Disambiguation (WSD) technique: WSD consists in recognizing and assigning the correct sense or Meaning of a given word. During the last years, many investigations on WSD have been done. These investigations can be mainly subdivided into three categories: Knowledge based [9][10], Supervised [11][12] and Unsupervised methods [13].

---

<sup>5</sup> based on preceding experiments  $K = 10$  was used

Knowledge-based approaches are based on the semantic resources such as ontologies. Supervised methods use manually annotated corpus for training classifiers. Unsupervised classifiers are trained on unannotated corpora to extract several groups of similar texts.

In our approach, to disambiguate the senses of the term and determine its descriptor in the context of the document, we use the language model approach. More precisely, to assess the relevance of a MeSH descriptor to this document, we estimate the probability that the MeSH descriptor would have been generated by language model of this document.

#### 4. OUR CONCEPTUAL INDEXING APPROACH

Our indexing methodology as schematized in Figure 1, consists of four main steps: (a) Pretreatment (b) term extraction (c) term weighing and (d) selection of descriptors. In the following, we describe the structure of MeSH vocabulary and then we detail the steps of our indexing method.

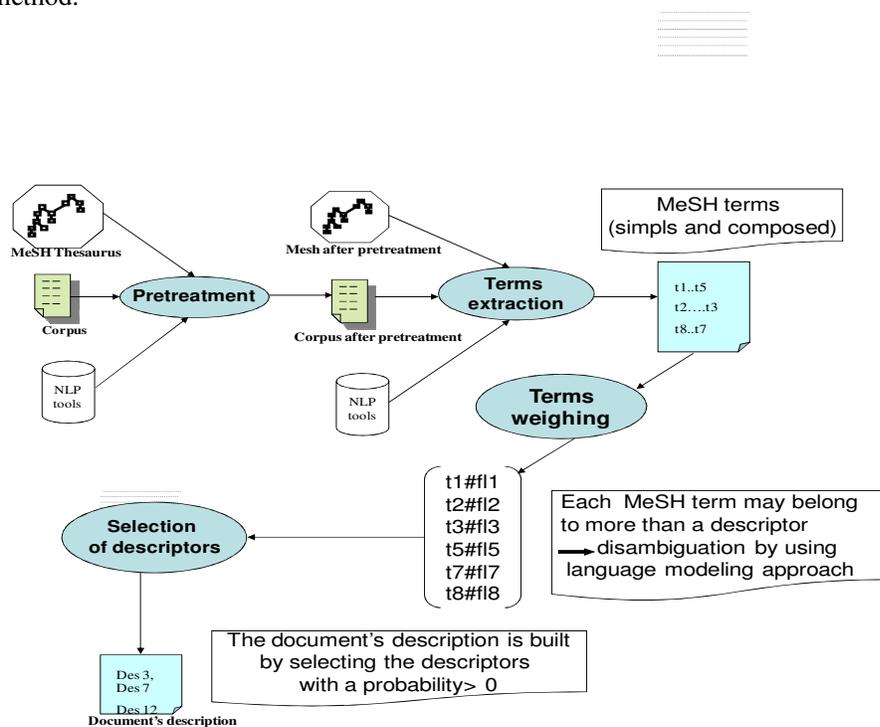


Figure 1. Architecture of our indexing approach

##### 4.1. MeSH thesaurus

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

- Each term can be either a simple or a composed term.
- A concept is viewed as a class of synonyms terms. The 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 related to the preferred concept by a relationship (broader, narrower).

Cardiomegaly [Descriptor]  
Cardiomegaly [Concept, Preferred]  
Cardiomegaly [Term, Preferred]  
Enlarged Heart [Term]  
Heart Enlargement [Term]  
Cardiac Hypertrophy [Concept, Narrower]  
Cardiac Hypertrophy [Term, Preferred]  
Heart Hypertrophy [Term]

Figure 2. Extrait of MeSH

As shown by figure 2, the descriptor “Cardiomegaly” consists of two concepts: “Cardiomegaly” and “Cardiac Hypertrophy”. Each concept has a preferred term, which is also said to be the name of the Concept.

For example, the concept “Cardiomegaly” has three terms “Cardiomegaly” (preferred term) , “Enlarged Heart” and “Heart Enlargement”.

As in the example above, the concept “Cardiac Hypertrophy” is narrower to than the preferred concept “Cardiomegaly”.

#### 4.2. Pretreatment

The first step is to split text into a set of sentences. We use the Tokeniser module of GATE [14] in order to split the document into tokens, such as numbers, punctuation, character and words. Then, the TreeTagger [30] 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 pre-treatment phase.

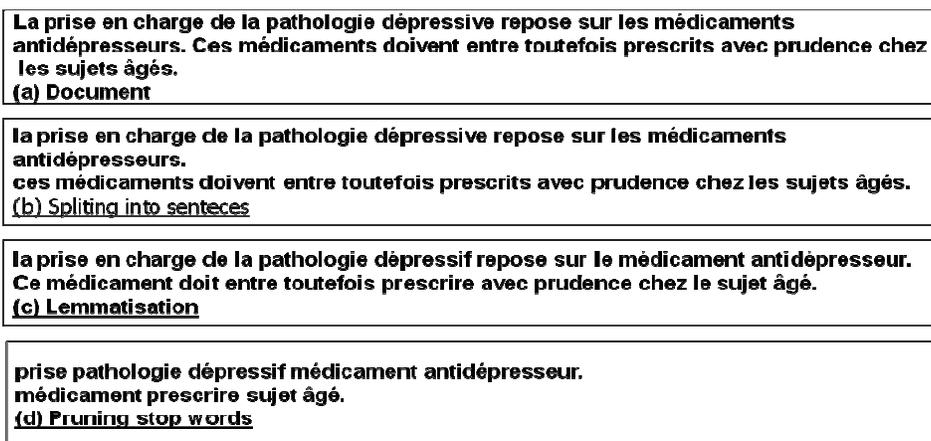


Figure 3. Pretreatment step

### 4.3. Term extraction

Automatic term extraction in textual corpora is a challenging task because terms are generally composed of multiple words. The current term extraction approaches can be categorized as (1) linguistic [15] [16], (2) statistical [17] [18] and (3) hybrid (linguistic and statistical) [19] approaches.

The linguistic approach allows defining, identifying and recognising terms looking at pure linguistic properties, using linguistic filtering techniques aiming to identify specific syntactic term patterns [20].

The statistical approaches are based on quantitative techniques such as collocation measure. A collocation, as defined by Choueka [21], is a sequence of adjacent words that frequently appear together.

The hybrid approaches takes into account both linguistic and statistical hints to recognise terms. As mentioned above, a term can be either simple or composed. To extract the simple term, we project the Mesh thesaurus on the document by applying a simple matching. More precisely, each lemmatized term in the document is matched with the canonical form or lemma of MeSH terms. To recognize the composed terms, we have chosen to use YateA [22]. YateA (Yet Another Term ExtrAator) is an hybrid term extractor developed in the project ALVIS. After text processing, YateA generates a file composed of two columns: the inflected form of the term and its frequency. For instance, as shown in figure 4 which describes the result of the term extraction process by using YateA, the term “exercice physique” occurs 6 times.

#	Inflected form	Frequency
	activité physique	16
	activité sportive	9
	exercice musculaire	8
	exercice physique	6
	effets bénéfiques	6
	g de glucides	5
	contrôle glycémique	5
	insuffisance coronaire	5
	index glycémique	4
	risque cardiovasculaire	4
	adaptation des doses	4
	glycémie capillaire	4
	sensibilité à l' insuline	4
	fréquence cardiaque	3
	hydrates de carbone	3
	acides gras libres	3
	autosurveillance glycémique	3
	patient dnid	3
	acides gras	3
	dernier repas	3
	profil lipidique	3
	activité physique régulière	3
	insuline rapide	3

Figure 4. An excerpt of the result of YaTeA

#### 4.4. Term weighing

Calculating term importance is a significant and fundamental step in information retrieval process and it is traditionally determined through term frequency (tf) and inverse document frequency (IDF).

In this paper we present a novel weighing technique by intuitively interpreting the conceptual information in the Mesh thesaurus. To determine term importance, this technique exploits its relationships in Mesh, rather than relying only on a statistical measure. Hence, the term weight is calculated by using its synonyms occurrences. Term significance is also captured using its location. More precisely, we assign an additional weight to the terms that are extracted from the title or the abstract of the document. To do so, we use two measures: the Content Structure Weight (CSW) and the Semantic Weight (SW).

##### 4.4.1. Content Structure Weight

We can notice that the frequency is not a main criterion to calculate the CSW of the term. Indeed, the CSW takes into account the term frequency in each part of the document rather than the whole document. For example, a term of the Title receives a higher importance (\*10) than to a term that appears in the Paragraphs (\*2). Table 1 shows the various coefficients used to weight the term locations. These coefficients were determined in an experimental way in [23].

Table 1. Weighing coefficients

Term location	Weight of the location
Title (T)	10
Keywords (K)	9
Abstract (A)	8
Paragraphs (P)	2

The CSW of the term  $t_i$  in a document  $d$  is given as follows:

$$CSW(t_i, d) = \frac{\sum_{A \in \{T, K, A, P\}} f(t_i, d, A) \times W_A}{\sum_{A \in \{T, K, A, P\}} f(t_i, d, A)} \quad (1)$$

Where:

–  $W_A$  is the weight of the location  $A$  (see Table 1),

–  $f(t_i, d, A)$  is the occurrence frequency of the term  $t_i$  in the document  $d$  at location  $A$ .

For example, the term *cancer* exists in the document  $d_{1683}$ : 1 time in the title, 2 times in the abstract and 9 times in the Paragraphs,

$$CSW(cancer, d_{1683}) = \frac{1 \times 10 + 2 \times 8 + 9 \times 2}{1 + 2 + 9}.$$

#### 4.4.2. Semantic Weight

The Semantic Weight of term  $t_i$  in the document  $d$  depends on its synonyms existing in the set of Candidate Terms ( $CT(d)$ ) generated by the term extraction step. To do so, we use the Synof function that associates for a given term  $t_i$ , its synonyms among the set of  $CT(d)$ .

Formally the measure SW is defined as follows:

$$SW(t_i, d) = \frac{\sum_{g \in Synof(t_i, CT(d))} f(g, d)}{|Synof(t_i, CT(d))|} \quad (2)$$

For a given term  $t_i$ , we have on the one hand its Content Structure Weight ( $CSW(t_i, d)$ ) and on the other its Semantic Weight ( $SW(t_i, d)$ ), its Local Weight ( $LW(t_i, d)$ ) is determined as follows:

$$LW(t_i, d) = \frac{CSW(t_i, d) + SW(t_i, d)}{2} \quad (3)$$

By examining the equation 3, we can notice that the terms (simple or composed) are weighted by the same way. Despite the several works dealing with the weighing of composed terms, there is so far no weighing technique shared by the community [24]. In our approach, we applied the weighing method proposed by [25]. According to [25], for a term  $t$  composed of  $n$  words, its frequency in a document depends on the frequency of the term itself, and the frequency of each sub-term. For this purpose, it proposes the measure  $cf$  which is defined as follows:

$$cf(t, d) = f(t, d) + \sum_{st \in subterms(t)} \frac{length(st)}{length(t)} \cdot f(st, d) \quad (4)$$

–  $f(t, d)$ : the occurrences number of  $t$  in the document  $d$ .

–  $length(t)$  represents the number of words in the term  $t$ .

–  $subterms(t)$  is the set of all possible terms MeSH which can be derived from  $t$ .

For example, if we consider a term “cancer of blood”, knowing that “cancer” is itself also a MeSH term, its frequency is computed as:

$$cf(\text{cancer of blood}, d) = f(\text{cancer of blood}, d) + \frac{1}{2} \cdot f(\text{cancer}, d)$$

Consequently, in an attempt to take into account the case of composed terms, we calculate the  $csw$  measure as follows:

$$CSW(t_i, d) = \frac{\sum_{A \in T, K, A, P} f(t_i, d, A) \times W_A}{\sum_{A \in T, K, A, P} (f(t_i, d, A))} + \sum_{st \in subterms(t_i)} \frac{length(st)}{length(t_i)} \cdot f(st, d) \quad (5)$$

Where:  $f(st, d)$  is the occurrences number of  $st$  in the document  $d$ .

#### 4.5. Selection of descriptors

A term MeSH may be located in different hierarchies at various levels of specificity, which reflects its ambiguity. As an illustration, figure 5 depicts the term “Pain”, which belongs to four

branches of three different hierarchies (descriptors) whose the most generic descriptors are: Nervous System Disease (C10); Pathological Conditions, Signs and Symptoms (C23); Psychological Phenomena and Processes (F02); Musculoskeletal and Neural Physiological Phenomena (G11).

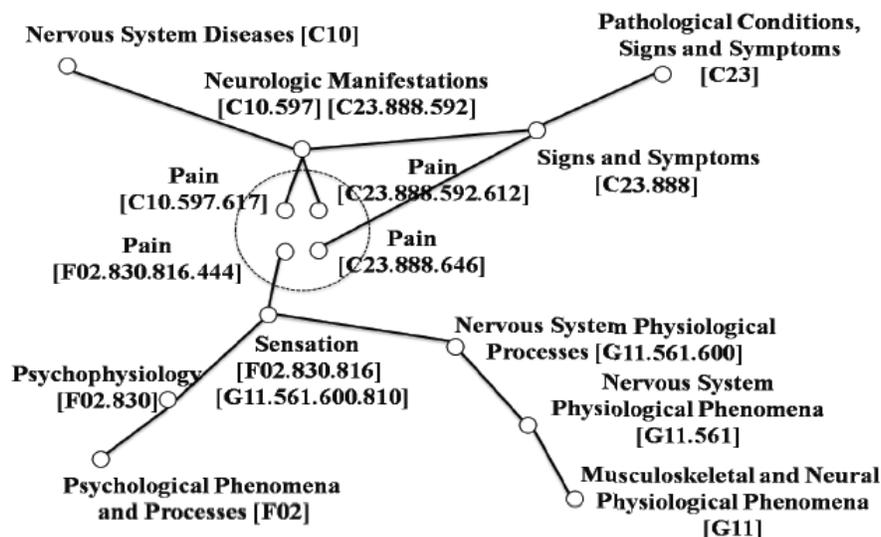


Figure 5. Term Pain in MeSH

In the last years, due to the amount of ambiguous terms and their various senses used in biomedical texts, term ambiguity resolution becomes a challenge for several researchers [26] [27] [28]. Differently from the proposed works in the literature, our method assign the appropriate descriptor related to a given term by using the language model approach.

For an ambiguous term, the task of WSD consists in answering the question: among its several senses, which is the best descriptor that can represent this term. The task of the WSD system is then to estimate, for each candidate descriptor MeSH, which is most likely to be the ideal concept. Hence, 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. Our basic assumption behind descriptor selection is that “for a given term  $t_i$  having several senses in the document  $d$ , the best descriptor is the one which has the highest probability to be generated by the model of this document”.

To do so, we estimated  $P(\text{des}|d)$  the probability of generating the descriptor  $\text{des}$  according to the document model. This probability is estimated by using the language model approach proposed by [29].

In the following, after a brief presentation of language model approach proposed by [29], we detail our WSD method.

#### 4.5.1. Language modeling approach

As is the case with the majority of language modeling approaches, [29] assumes that the user has a reasonable ideal of the terms that are likely to appear in the ideal document that can satisfy his information need, and the query terms the user chooses can distinguish the ideal document from the rest of the collection. The query is then generated as the piece of text representative of the ideal document. For each document  $d$  in the collection, [29] calculates the probability of generating the user query by using this equation:

$$P(d|Q) = \prod_{t_i \in Q} ((1 - \lambda)P(t_i|d) + \lambda P(t_i|D)) \quad (7)$$

Where:

–  $P(t_i|D)$  represents the probability of generating  $t_i$  from corpus  $D$ .

$$P(t_i|D) = \frac{df(t_i, D)}{\sum_{t_j \in D} df(t_j, D)} \quad (8)$$

$df(t_j, D)$  the number of documents which term  $t_j$  is occurs in.

–  $P(t_i|d)$  represents the probability of generating  $t_i$  from a document  $d$ .

$$P(t_i, d) = \frac{tf(t_i, d)}{|d|} \quad (9)$$

$tf(t_i, d)$  is the frequency of the term  $t_i$  in the document  $d$ .

Finally, documents are ranked according to their estimated degree or probability of usefulness for the user.

#### 4.5.2. WSD using language modelling

In our WSD approach, to determine for an ambiguous term, its best descriptor, we have adapted the language model of [29] by substituting the query by the Mesh descriptor. Thus, we infer a language model for each document and rank Mesh descriptors according to their probability of producing each one given this model. We would like to estimate  $P(\text{des}|d)$ , the probability of generation a Mesh descriptor  $\text{des}$  given the language model of document  $d$ .

For a collection ( $D$ ), a document ( $d$ ) and a MeSH descriptor ( $\text{des}$ ) composed of  $n$  concepts, the probability  $P(\text{des}|d)$  is done by :

$$P(\text{des}_k|d) = \prod_{c_j \in \text{relatedtoDes}(\text{des}_k, d)} ((1 - \lambda)P(c_j|d) + \lambda P(c_j|D)) \quad (10)$$

RelatedtoDes (respectively RelatedtoCon) is the function that associates for a given descriptor  $\text{des}$  (respectively concept  $\text{con}$ ) and a document  $d$ , the concepts (respectively terms) MeSH which are related to  $\text{des}$  (respectively  $\text{con}$ ) in  $d$ .

In the equation 10, we need to estimate two probabilities:

- $P(c_j|D)$ : the probability of observing the concept  $c_j$  in the collection  $D$ .

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

$$f(c_j, D) = \sum_{t_i \in \text{relatedtoCon}(c_j, D)} df(t_i, D)$$

$df(t, D)$ :  $df$  (document frequency) is the number of documents which term  $t$  occurs in  $D$ .

- $P(c_j|d)$ : the probability of observing a concept  $c_j$  in a document  $d$ :

$$P(c_j|d) = \frac{f(c_j, d)}{|\text{concepts}(d)|}$$

$$f(c_j, d) = \sum_{t_i \in \text{relatedtoCon}(c_j, d)} LW(t_i, d)$$

$LW(t, d)$  is determined by using the equation 3.

Based on this approach, to assign the appropriate sense (Best Descriptor (BD)) related to an ambiguous term ( $t_i$ ) in the context of document ( $d_j$ ), we must go through these steps:

### 1. Compute the descriptor relevance score:

Let  $senses_{d_j}^i = \{des_{d_j}^{i1}, des_{d_j}^{i2}, des_{d_j}^{i3} \dots des_{d_j}^{in}\}$ : the set of descriptors MeSH that can represent the term  $t_i$  in the document  $d_j$ . For each descriptor  $des_k$  existing in this set, we need to measure its ability to represent the term ( $t_i$ ) in the document ( $d_j$ ). To do so, we calculate  $P(des_k | d_j)$  (see equation 10).

### 2. Selection of the best descriptor:

The best descriptor (BD) to retain is the one which maximizes  $P(des_k | d_j)$ :

$$BD(t_i, d_j) = \max_{des_k \in senses_{d_j}^i} p(des_k | d_j) \quad (13)$$

Finally, the Semantic Index (SI) of document  $d$ , is generated as follows:

$$SI(d_j) = \bigcup_{t_i \in CT(d_j)} BD(t_i, d_j) \quad (14)$$

## 5. EXPERIMENTAL EVALUATION

In order to assess the feasibility of our indexing approach, we have carried out a BIOmedical Indexing System (BIOINSY). Our indexing system was evaluated by intensive experiments. In order to make clear these experiments, we first present shortly the training data sets. After that,

we describe the experimental process and the techniques used for validation. Finally, we discuss the obtained results.

## 5.1. Training data sets

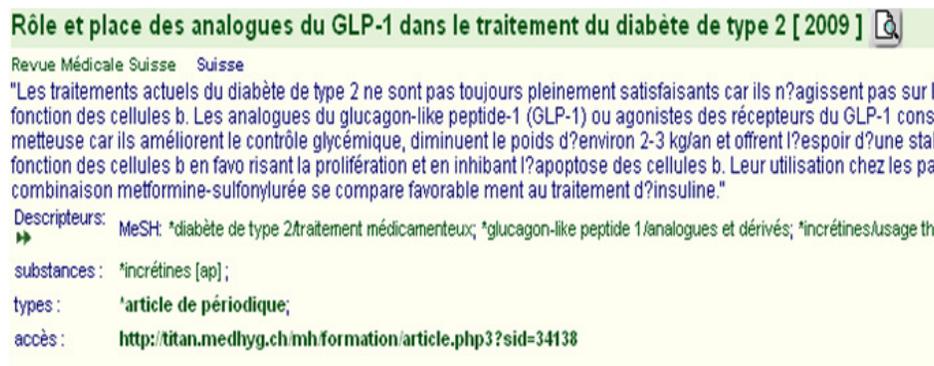
To evaluate our indexing approach we built a corpus consisting of 500 scientific articles selected from CISMEF<sup>6</sup>. In this collection, we tried to have a diversified base in terms of content. This corpus was carried out with a set of medical keywords (cancer, diabetes, pregnancy, fever...) used on the search engine of CISMEF. Analysis of this corpus revealed about 716,000 words. Each document of this corpus has been manually indexed by five professional indexers in the CISMEF team in order to provide a description called “notice”.

Figure 6 shows an example of notice for resource “diabete de type 2”.

A notice is mainly composed of:

- General description: title, authors, abstract...
- Specific description: MeSH descriptors that describe article content.

In this evaluation, the notice (manual indexing) is used as a reference. In fact, 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).



**Rôle et place des analogues du GLP-1 dans le traitement du diabète de type 2 [2009]** 

Revue Médicale Suisse Suisse

"Les traitements actuels du diabète de type 2 ne sont pas toujours pleinement satisfaisants car ils n'agissent pas sur la fonction des cellules b. Les analogues du glucagon-like peptide-1 (GLP-1) ou agonistes des récepteurs du GLP-1 constituent une alternative intéressante car ils améliorent le contrôle glycémique, diminuent le poids d'environ 2-3 kg/an et offrent l'espoir d'une stabilisation de la fonction des cellules b en favorisant la prolifération et en inhibant l'apoptose des cellules b. Leur utilisation chez les patients avec une combinaison metformine-sulfonylurée se compare favorablement au traitement de l'insuline."

Descripteurs: MeSH: \*diabète de type 2/traitement médicamenteux; \*glucagon-like peptide 1/analogues et dérivés; \*incrétines/usage thérapeutique

substances: \*incrétines [ap];

types: \*article de périodique;

accès: <http://titan.medhyg.ch/mhformation/article.php3?sid=34138>

Figure 6. CISMeF notice for resource “Diabete de type 2”

## 5.2. Experimental process

In this experimental process, three measures have been used: precision, recall and F-measure. Precision corresponds to the number of indexing concepts correctly retrieved over the total number of retrieved concepts.

Recall corresponds to the number of indexing concepts correctly retrieved over the total number of expected concepts.

F-measure combines both precision and recall with single measure.

<sup>6</sup> Catalogue et Index des Sites Médicaux Francophones

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

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

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

Where:

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

### 5.3. Experimental results

To evaluate the effectiveness of our weighing method, we carried out three sets of experiments:

- Experiment 1: classical term weighing: the term is weighed by using the measures (tf/idf).
- Experiment 2: semantic term weighing: the term is weighed by using the measures CSW (see equation and SW (see equations 1 and 2).
- Experiment 3: composed term weighing: in this case the measures CSW is calculated by using the equation 5.

Table 2 shows the precision (P) and the recall (R) obtained by our system BIOINSY at fixed ranks 1 through 10 in each case cited above.

Table 2. Precision and recall generated by BIOINSY

Rank	Experiment1(P/R)	Experiment 2(P/R)	Experiment 3(P/R)
1	17,96/20,67	19,87/22,04	21,25/24,96
4	31,19/18,81	51,98/20,13	45,15/20,13
10	48,96/9,14	69,85/8,29	68,72/9,08

In order to make clear these experimental results, we propose the figure 7 which presents the F-measure value generated by BIOINSY at ranks 1, 4 and 10.

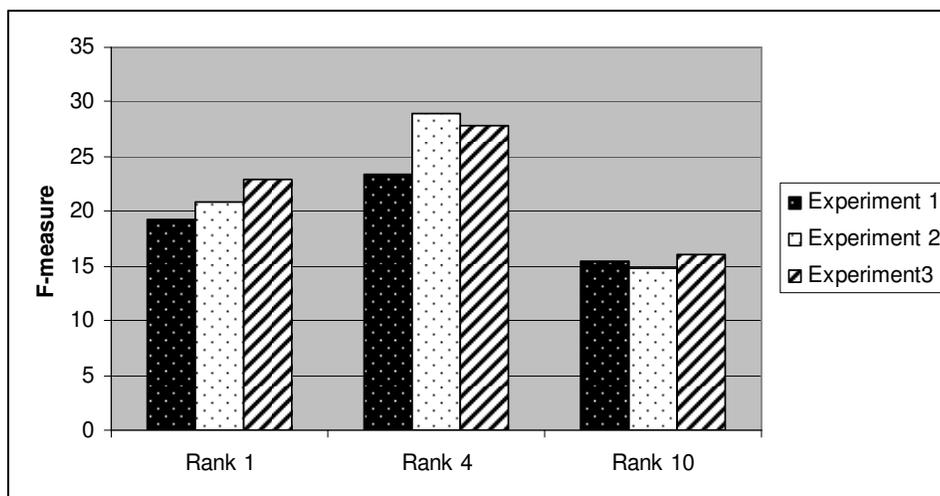


Figure 7. The F-measure value generated by BIOINSY at ranks 1, 4 and 10 in three sets of experiments

As shown in figure7, at all ranks, our semantic based weighing approach (case 2) is really significant compared to the classical term weighing method (case 1). For instance by using YateA, at rank 4, for recall BIOINSY displayed 31, 19% in the case 1 and 51, 98% in the case 2. The obtained results confirm the well interest to integrate the composed term technique in the process weighing. In most of cases (rank 1 and 10), the experiment 3 shows the best F-measure value. The experimental results highlight the performance of weighing method proposed in this paper.

## 6. COMPARISON OF BIOINSY WITH OTHERS TOOLS

To prove the effectiveness of our indexing method, we compared the BIOINSY to other medical indexing systems. We evaluate the performance of six indexing systems (MetaMap, EAGL, KNN, MTI and BIOINSY) in terms of generating the manual MeSH annotations. For this evaluation, we used the same corpus<sup>7</sup> used by [33] composed of 1000 random MEDLINE citations.

Table 3 shows the results generated by indexing systems using the title of a 1000 random MEDLINE citations.

Table 3. Precision and MAP(Mean Average Precision) generated by the indexing systems using title as input

Indexing system	Precision at rank10 (P@10)	MAP
MTI	0.18	0,16
MetaMap	0.17	0,14
EAGL	0.18	0,17
KNN	0.43	0,47
<b>BIOINSY</b>	<b>0,39</b>	<b>0,43</b>

<sup>7</sup> The corpus can be downloaded in ([http://www.ebi.ac.uk/~triesch/meshup/testset\\_v1.xml](http://www.ebi.ac.uk/~triesch/meshup/testset_v1.xml)).

Table 4 shows the results generated by indexing systems with the title and abstract of a 1000 random MEDLINE citations.

Table 4. Precision and MAP(Mean Average Precision) generated by the indexing systems using title and abstract as input

Indexing system	Precision at rank10 (P@10)	MAP
MTI	0.32	0.25
MetaMap	0.19	0.16
EAGL	0.21	0.19
KNN	0.45	0.50
<b>BIOINSY</b>	<b>0,30</b>	<b>0,26</b>

Figure 8 illustrates the obtained results by the five indexing systems on the 1000 random MEDLINE citations.

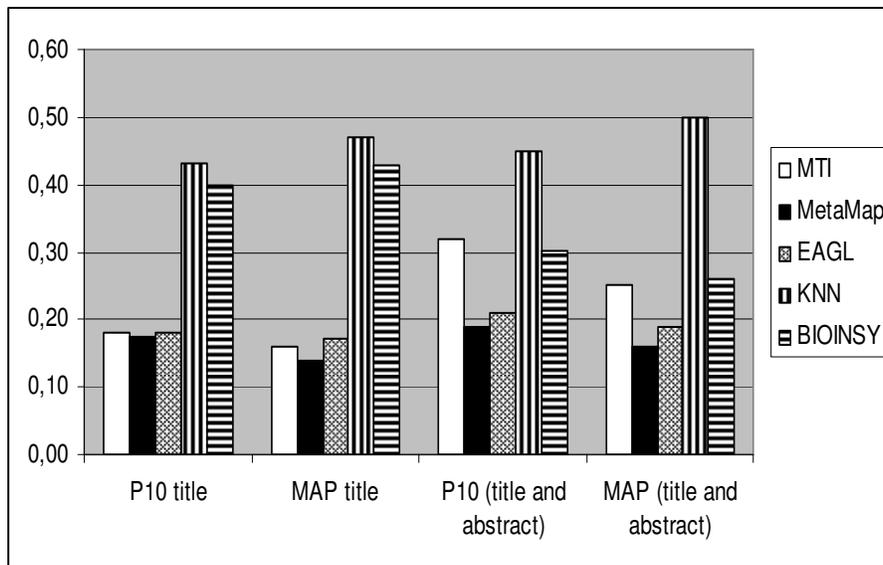


Figure 8. Experimental results generated by the five indexing systems

Our system BIOINSY serves as the baseline against which the other systems are compared. Both indexing systems MetaMap and EAGL perform worse than BIOINSY in all metrics. Indeed, MetaMap performs similarly to or slightly worse than EAGL when presented with the title only or both title and abstract of the citation to index.

MTI performs worse than BIOINSY when the title was available for indexing. For example, when title used as input, the value of P10 generated by MTI is equal to 0,18. Concerning BIOINSY, it generates 0,39 as value of P10. When using title and abstract, MTI performs similarly to or slightly better than BIOINSY in terms of MAP and P10.

By using title as input, KNN and BIOINSY echoed very similar performance. Given the title and abstract of a citation, KNN shows the best results in all metrics.

The obtained results confirm the well interest to use the language modeling approach in the conceptual indexing process.

## 7. CONCEPTUAL RETRIEVER

In this section, we try to answer the following question: Can our conceptual indexing approach (described and evaluated above) improve the information retrieval process. The overview of this section is as follows. In subsection 7.1 we will present the test collection. In subsection 7.2 we will describe the experimental setup. In subsection 7.3, the experimental results will be analysed and discussed.

### 7.1. Test collection

To evaluate the retrieval effectiveness based on our conceptual indexing method, we use the ImageCLEF<sup>8</sup>med 2007 collection. Started from 2004, the ImageCLEFmed (medical retrieval task) aims at evaluating the performance of medical information systems, which retrieve medical information from a mono or multilingual image collection.

This corpus [31] is based on a dataset containing images from the Casimage, MIR, PEIR, PathoPIC, CORI, myPACS and Endoscopic collections. For each image of this corpus, a textual description called diagnosis is attributed. An overview of databases used in ImageCLEFmed 2007 is shown in Table 5.

Table 5. details of the ImageCLEFmed 2007 collection

Collection Name	Cases	Images	Annotations	Annotations by Language
Casimage	2076	8725	2076	French – 1899, English – 177
MIR	407	1177	407	English – 407
PEIR	32319	32319	32319	English – 32319
PathoPIC	7805	7805	15610	German – 7805, English – 7805
myPACS	3577	15140	3577	English – 3577
Endoscopic	1496	1496	1496	English – 1496
Total	47680	66662	55485	French – 1899, English – 45781, German – 7805

### 7.2. Experimental setup

The ImageCLEF data contains the qrels file (TREC format) which specifies the set of relevant images to a given query. In our indexing method we are interested by the textual document. Hence, to evaluate our approach we assume that “If a query is relevant to an image then it is also relevant to its textual description (diagnosis)”.

This evaluation process is structured around the following steps:

<sup>8</sup> CLEF (Cross Language Evaluation Forum)

- **Indexing of diagnosis and queries:**

The indexing process is carried out on the diagnosis and queries. Thus, documents and eventually queries are expanded with descriptors identified by our conceptual indexing method.

The documents and the queries are presented as follows:

$$\vec{d}_j = (w_j^1, w_j^2, \dots, w_j^n)$$

- **Calculation of Retrieval Status Value (RSV (q, d)):**

The relevance score of the document  $d_j$  with respect to the query  $q$  is given by:

$$RSV(q, d_j) = \sum_{des \in Q} TF_j(des) \times IDF(des)$$

Where:

- $TF_j$ : the normalized term frequency of the current descriptor in document  $d_j$ .
- $IDF$ : the normalized inverse document frequency of the current descriptor in the collection.

### 7.3. Results and discussion

In order to assess the validity of our method, we compared the results of our indexing system BIOINSY to official runs in medical retrieval task 2007.

Table 6 summarizes the results obtained by the participants in medical retrieval task 2007.

Table 6: The comparison of our system with official runs participated in ImageCLEF9med 2007

Run	P@5	MAP
LIG-MRIM-LIGMU (the best)	0.44	0,32
OHSU (the second)	0.42	0,27
IPAL4 (median)	0.39	0,27
miracleTxtFRT (median)	0.43	0,17
IRIT RunMed1 (the worst)	0,05	0,04
<b>Our system</b>	<b>0,38</b>	<b>0,24</b>

By examining the table 6, we can note that the results generated by our system close to those of the best run (LIG-MRIM-LIGMU). Thus, we conclude that our conceptual indexing approach proposed in this paper would significantly improve the biomedical IR performance.

## 8. CONCLUSION

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 principally of three main steps. At the first step (Term extraction), being given an article, MeSH thesaurus and the NLP tools, the system

<sup>9</sup> CLEF (Cross Language Evaluation Forum)

BIOINSY extracts two sets: the first is the article's lemma, and the second is the list of lemma existing in the MeSH thesaurus. After that, these sets are used in order to extract the Mesh terms existing in the document. At step 2, these extracted terms are weighed by using the measures CSW and SW that intuitively interprets MeSH conceptual information to calculate the term importance. The step 3 aims to recognize the MeSH descriptors that represent the document by using the language model.

In order to assess its feasibility, our indexing approach was experimented on through training data sets containing 500 medical articles. An experimental evaluation and comparison of BIOINSY with others indexing tools confirms the well interest to use the language modelling approach in the conceptual indexing process. The performance of our approach is also close to the average of official runs in CLEFmedical retrieval task 2007 and is comparable to the best run.

Our future work aims at incorporating a kind of semantic smoothing into the language modeling approach. We also plan to use several semantic resources in the indexing process. We believe that multi-terminology based indexing approach can enhance the IR performance.

## REFERENCES

- [1] Pouliquen, B.: Indexation de textes médicaux par indexation de concepts, et ses utilisations. PhD thesis, Universit Rennes 1 (2002)
- [2] Neveol, A.: Automatisation des taches documentaires dans un catalogue de santé en ligne. PhD thesis, Institut National des Sciences Appliquées de Rouen (2005)
- [3] Dinh, D., Tamine, L.: Biomedical concept extraction based on combining the content-based and word order similarities. In: SAC. (2011) 1159–1163
- [4] A.Aronson, J.Mork, C.S., W.Rogers: The nlm indexing initiative's medical text indexer. In: Medinfo. (2004)
- [5] A.Aronson: Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In: AMIA. (2001) 17–21
- [6] W.Kim, A.Aronson, W.: Automatic mesh term assignment and quality assessment. In: AMIA. (2001)
- [7] P.Lenoir, R.Michel, C., G.Chales: Réalisation, développement et maintenance de la base de données a.d.m. In: Médecine informatique. (1981)
- [8] Zakos, J., Verma, B.: Concept-based term weighting for web information retrieval. Computational Intelligence and Multimedia Applications, International Conference on 0 (2005) 173–178
- [9] Lesk, M.: Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice creamcone. In: SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation, New York, NY, USA, ACM (1986) 24–26
- [10] Mihalcea, R.: Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. HLT '05, Stroudsburg, PA, USA, Association for Computational Linguistics (2005) 411–418
- [11] Lee, Y.K., Ng, H.T., Chia, T.K.: Supervised word sense disambiguation with support vector machines and multiple knowledge sources. In: Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text. (2004) 137–140 12
- [12] Liu, H., Teller, V., Friedman, C.: A multi-aspect comparison study of supervised word sense disambiguation. J Am Med Inform Assoc (11) 320–31
- [13] Yarowsky, D.: Unsupervised word sense disambiguation rivalling supervised methods. In: In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. (1995) 189–196
- [14] Cunningham, M., Maynard, D., Bontcheva, K., V.Tablan: Gate: A framework and graphical development environment for robust NLP tools and applications. ACL (2002)

- [15] Bourigault, D.: Extraction et structuration automatique de terminologie pour l'aide l'acquisition des connaissances partir de textes. In: 9me congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'94), Paris. (1994) 397–408
- [16] Jacquemin, C., Royaut, J.: Retrieving terms and their variants in a lexicalized unification-based framework. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. (1994) 132–141
- [17] drouin, P., LADOUCEUR, J.: (L'identification automatique de descripteurs complexes dans des textes de spécialité)
- [18] Oueslati, R.: Aide l'acquisition de connaissances partir de corpus. PHD thesis, Université Louis Pasteur (1999)
- [19] daille, B.: Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques. PhD thesis, Université Paris 7 (1994)
- [20] Paziienza, M., Pennacchiotti, M., Zanzotto, F.: Terminology extraction: An analysis of linguistic and statistical approaches. In Sirmakessis, S., ed.: Knowledge Mining Series: Studies in Fuzziness and Soft Computing. Springer Verlag (2005)
- [21] Choueka, Y.: (Looking for needles in a haystack or locating interesting collocational expressions in a large textual database)
- [22] Aubin, S., Hamon, T.: Improving term extraction with terminological resources. In: Advances in Natural Language Processing. Volume 4139 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2006) 380–387
- [23] Gamet, J.: Indexation de pages web. Report of dea, university de Nantes (1998)
- [24] Baziz, M., Boughanem, M., Aussenac-Gilles, N., Chrisment, C.: Semantic cores for representing documents in ir. In: Proceedings of the 2005 ACM symposium on Applied computing. SAC '05, ACM (2005) 1011–1017
- [25] Baziz, M.: Indexation conceptuelle guide par ontologie pour la recherche d'information. PhD thesis, Univ. of Paul sabatier (2006)
- [26] Andreopoulos, B., Alexopoulou, D., Schroeder, M.: Word sense disambiguation in biomedical ontologies with term cooccurrence analysis and document clustering. IJDMB 2 (2008) 193–215
- [27] Stevenson, M., Guo, Y., Gaizauskas, R., Martinez, D.: Knowledge sources for word sense disambiguation of biomedical text. In: BioNLP '08: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, Association for Computational Linguistics (2008) 80–87
- [28] Dinh, B., Tamine, L.: Sense-based biomedical indexing and retrieval. In: NLDB. (2011) 24–35
- [29] Hiemstra, D.: Using Language Models for Information Retrieval. PhD thesis, University of Twente (2001)
- [30] H.Schmid: Probabilistic part-of-speech tagging using decision trees. In: International Conference on New Methods in Language Processing, Manchester (1994)
- [31] Miller, H., Deselaers, T., Deserno, T., Clough, P., Kim, E., Hersh, W.: Overview of the imageclefmed 2006 medical retrieval and annotation tasks. In: In: CLEF 2006 Proceedings. Lecture Notes in Computer Science (2006) 595–608
- [32] Ruch, P. Automatic assignment of biomedical categories: toward a generic approach. Bioinformatics, 22, (2006) 658–664.
- [33] Trieschnigg D., Pezik P., Lee V., Kraaij W., de Jong F., and Rebholz-Schuhmann D. MeSH Up: Effective MeSH Text Classification and Improved Document Retrieval. Bioinformatics 25(11), (2009), 1412–1418.