Solving Concept mismatch through Bayesian Framework by Extending UMLS Meta-Thesaurus

Karam Abdulahhad, Jean-Pierre Chevallet, Catherine Berrut

Solving Concept mismatch through Bayesian Framework by Extending UMLS Meta-Thesaurus

Karam ABDULAHHAD* — **Jean-Pierre CHEVALLET**** — **Catherine BERRUT***

** UJF-Grenoble 1, ** UPMF-Grenoble 2, LIG laboratory, MRIM group*

{karam.abdulahhad,jean-pierre.chevallet,catherine.berrut}@imag.fr

ABSTRACT. *Most Information retrieval systems are based on exact term matching. Though many researches address the "term mismatch" problem. This problem arises when different terms express the same meaning, in multilingual formulation of the query/documents, or when using expert terms either in the document or in the query. All these problems need a particular analysis that fills the gap between the document information and the user. In this work, we propose a solution by the enrichment of a meta-thesaurus. We propose to exploit the thesaurus relations between concepts but also to enrich them through the analysis of the terms associated to concepts. The matching is performed using concept derivation through a Bayesian network. A validation of our proposal is made on the test collection ImageCLEFMed 2005 and the resource UMLS 2005.*

RÉSUMÉ. *La plupart des systèmes de Recherche d'Information sont basés sur la correspondance exacte entre termes, même si de nombreuses recherches portent sur le problème de la correspondance entre variantes de termes issus de mots synonymes, d'une formulation multilingue, ou sur l'utilisation de termes experts très précis. Résoudre ces problèmes nécessite une analyse particulière pour combler l'écart entre l'information contenue dans les documents et le besoin de l'utilisateur. Dans ce travail, nous proposons une solution par enrichissement d'un métathésaurus. Nous proposons d'exploiter les relations entre les concepts du thesaurus, mais aussi de les enrichir à travers l'analyse des termes associés aux concepts. La correspondance est effectuée en utilisant une dérivation de concepts à travers un réseau Bayésien. Une validation de notre proposition est réalisée sur la collection de test ImageCLEFmed 2005 avec l'aide de la ressource UMLS 2005.*

KEYWORDS: *term mismatch, concept mismatch, Bayesian matching, conceptual indexing*

MOTS-CLÉS : *variation terminologique, correspondance Bayésien, indexation conceptuelle*

1. Introduction

Information Retrieval System (IRS) is a mediator between a user and a corpus (set of documents). IRSs are used to retrieve documents that may contain relevant information. IRSs are based on dedicated models.

Classic IR models e.g. Boolean, Vector Space, Probabilistic, Bayesian (Turtle *et al.,* 1990) (Turtle *et al.,* 1991), etc. use the "bag of words" paradigm to represent documents and queries. They calculate how much a document matches a query by word intersection between document and query. Therefore, they cannot reach relevant documents that don't share any word with the query.

The question here is: '*Do intersection-based IRSs are suitable and sufficient for all types of users and in all domains of search?*'

All IRSs based on intersection to compute a matching between a document and a query, suffer from "**element mismatch**" problem (Crestani, 2000) (Baziz, 2005) (Maisonnasse, 2008) (Chevallet, 2009).

Indexing element are used to represent the content of documents and queries. So an element may be a word, a term, or a concept. A *term* is a noun phrase that has a unique meaning in a specific domain (e.g. medical domain) and that belongs to a terminology (Baziz, 2005) (Chevallet, 2009). A *concept* could be defined as "Human understandable unique abstract notions independent from any direct material support, independent from any language or information representation, and used to organize perception and knowledge" (Chevallet *et al.,* 2007). Practically, a concept is represented by an identifier from an external resource. Each concept is associated to a set of terms that describe it (Baziz, 2005) (Chevallet, 2009).

When using words as indexing elements, IRS suffers from **word mismatch** problem. This problem appears when using different words to represent the same meaning, e.g. (atrial, auricular), (apartment, flat), etc. For example, without an external resource, that links synonymic words, the system cannot retrieve a document containing '*apartment*' as a response to a query containing '*flat*'.

Some researchers use terms instead of words as indexing elements. They suppose that a phrase is more precise than individual words (Baziz, 2005). By using terms as indexing elements, IRS still suffers from a **term mismatch** problem. This problem appears when users write a query by using terms and document's author uses a different terms to express the same meaning. For example, the following two terms '*Skin Cancer*' and '*melanoma*' have a close meaning. Term mismatch is also related to term variation like '*air pollution*' vs. '*pollution of the air*'. In addition, in less than 20% of cases, two people use the same term to describe the same meaning (Crestani, 2000). Consequently, we can formulate the word/term mismatch problem as follows: "How often have you tried to look up something in an index and failed to find what you were looking for because the words or phrases you looked for were different from those used in the material you needed to find?" (Woods, 1997).

To solve the term mismatch problem, many researchers propose to use concepts as indexing elements. This solves only a part of the problem when different terms correspond to the same concept. For example, the two terms "*Atrial Fibrillation*" and "*Auricular Fibrillation*" correspond to the same concept "*C0004238*" in UMLS [1]. However, let's consider two terms associated to two different concepts having a semantic relation. If a system is unable to exploit this semantic link then it is unable to retrieve a document indexed with one concept with a query containing the other. By example, the two terms "*B-Cell*" and "*Lymphocyte*" correspond to the two concepts "*C0004561*" and "*C0024264*" respectively, and there is a relation of type "*isa*" between the two. Here again we are facing a similar mismatch problem but at conceptual level "**concept mismatch**".

From the previous discussion, it seems clear that the solution of element mismatch problem relies on an external resource (Baziz, 2005) (Chevallet *et al.*, 2007) (Maisonnasse, 2008) (Chevallet, 2009), which is crucial for: **1)** using concepts for indexing, **2)** linking different terms which have the same meaning, and **3)** exploiting relations between concepts.

We present in the next section some works aiming to solve term and concept mismatch problems by exploiting relations between concepts and by using a structure like a Bayesian Network (Le, 2009).

This paper is structured as follows: in section 2, we present existing solutions to element mismatch problem. In section 3, we explain our proposed model. In section 4, the actual context of model validation is showen. The experiments that we have done are then presented. We conclude in section 5.

2. Related works

Crestani (Crestani, 2000), presents three approaches to solve term mismatch [2] problem:

1) Dimensionally reduction: it reduces term space, so reducing the chance that a query and a document use different terms to represent the same meaning.

2) Query expansion: it considers the query as a tentative definition of user need, then by applying some techniques, the meaning of query could be expanded by adding other terms.

3) Imaging (Crestani *et al.*, 1995): it solves term mismatch problem by influencing document term weights through similar terms not seen in the document. In this approach, term space and queries are not changed.

1. Unified Medical Language System. It is a meta-thesaurus in medical domain. http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=nlmumls

2. maybe the word "term" here doesn't correspond exactly to the definition of term that we have presented before, but the principle is same

According to Crestani (Crestani, 2000), none of the previous approaches completly solve term mismatch problem, and each of them has its drawback.

Crestani also proposes a solution to term mismatch problem by exploiting the similarity of non-matching terms at retrieval time. He supposes that there is a measure able to estimate the semantic similarity between pairs of terms.

$$\forall (t_i, t_j) \in T, 0 \leq Sim(t_i, t_j) \leq 1 \quad [1]$$

Then he exploits this measure at retrieval time by applying one of the following two formulas (considering the point of view of the query q):

$$RSV_{max(q \triangleright d)}(d, q) = \sum_{t \in q} Sim(t, t^*) w_d(t^*) w_q(t) \quad [2]$$

$$RSV_{tot(q \triangleright d)}(d, q) = \sum_{t_k \in q} \left(\sum_{t_j \in d} Sim(t_k, t_j) w_d(t_j) \right) w_q(t_k) \quad [3]$$

Where:
$q \triangleright d$ means, we start from query's terms and compare them with document's terms
$t^* \in T$ is a document's term that produces the maximum similarity with t, a query term.
$w_d(t^*)$ is the weight of term t^* in the document d.
$w_q(t)$ is the weight of term t in the query q.
Considering the point of view of the document d:

$$RSV_{max(d \triangleright q)}(d, q) = \sum_{t \in d} Sim(t, t^*) w_d(t) w_q(t^*) \quad [4]$$

$$RSV_{tot(d \triangleright q)}(d, q) = \sum_{t_j \in d} \left(\sum_{t_k \in q} Sim(t_k, t_j) w_q(t_k) \right) w_d(t_j) \quad [5]$$

The previous formulas take into account the non-matching terms of a query when computing RSV between a document d and a query q. It does not change the term space nor expand the query.

However, Crestani's proposed solution has some drawbacks, among them: **1)** the difficulty to find effective similarity measure, **2)** the high cost of computing the measure for large term space.

Chevallet and al (Chevallet *et al.*, 2007) proposes another solution to term mismatch problem by using concepts instead of terms but they do not exploit concept links.

Diem Le in her PhD thesis (Le, 2009), proposes a solution to term and concept mismatch problems by using concepts as indexing elements and using semantic relations during the matching process.

This model consists of three main components $RIRBRE\left(\Upsilon,\Psi,\Delta\right)$ [3]:
1) The external resource Υ: it consists of terms, concepts and semantic relations between concepts.
2) Bayesian Network $\Psi\left(N,L,P\right)$: it represents the content of documents, queries and relationships between them. It consists of two components:
Nodes N: documents, query and the concepts that represent the content of documents and query.
Links L: links between documents and their concepts, links between a query and its concepts and links between the concepts of documents and the concepts of query. The last type of links represents the relationships between concepts.
3) Matching function Δ: $\Delta\left(d,q\right)=P\left(q|d\right)=bel\left(q\right)$.
The matching value is calculated as follows:
a) Start by choosing one document $d\in D$, forcing $P\left(d\right)=1$.
b) Calculate the conditional probability of concept nodes:

$$bel\left(c_i\right)=\begin{cases} w\left(c_i,d\right) & \text{if } c_i \text{ is one of the concepts of document } d \\ bel\left(c_h\right)\times sim\left(c_h,c_i\right) & \text{if } c_i \text{ is one of the concepts of query } q \text{ and } c_h \text{ is the most important concept in document } d \text{ which has a relation with } c_i \\ 0 & \text{else} \end{cases} \quad [6]$$

Where: Sim calculates the similarity between two concepts.
c) Now to calculate $bel\left(q\right)$:

$$\Delta\left(d,q\right)=P\left(q|d\right)=bel_{wsum}\left(q\right)=\frac{\sum_{c_i\in q}w\left(c_i,q\right)\times bel\left(c_i\right)}{\sum_{c_j\in q}w\left(c_j,q\right)} \quad [7]$$

The RIRBRE model exploits semantic relations between concepts founded in an external resource to compute the matching. RIRBRE has been tested in the medical domain using UMLS meta-thesaurus as an external resource. However, external resources are *incomplete*. This is the case for UMLS, even if it is the largest resources available for medical applications. Several studies show that many concepts and relations are missing in UMLS (Bodenreider *et al.*, 1998) (Bodenreider *et al.*, 2001), and there exists proposals to compensate this incompleteness: Bodenreider and al (Bodenreider *et al.*, 2001) postulates that terms with adjectival modifiers are potential hyponyms. They propose to remove the modifiers from a term t_1 to get another term t_2 in a relation of type hyponym to t_1 (t_1 is hyponym of t_2).

In our side, we have experienced the incompleteness of UMLS through some statistics. We have computed the number of concept pairs whom terms share at least on word and does not have any semantic relation. For example, we found five concepts containing the word "*spirochaete*". However, among the 20 potential relations between these concepts none are founded in UMLS. In Table 1, we present the descrepancy between existing semantic relations and potential relations based on word sharing.

3. IR Model based on Extended Bayesian Network using External Resource.

Table 1. *Statistics from UMLS*

word	# concepts	# all concepts pairs	# pairs with relation
device	86,985	7,566,303,240	161,660
activity	22,395	501,513,630	380,052
sedum	98	9,506	122
spirocheate	5	20	0

3. A solution to term and concept mismatch

We propose in this work to enrich the external resource by adding relations between concepts, and to profit from the new relations beside the old ones. We make the hypothesis that by adding simple relations between concepts and using them in the retrieval process, more relevant documents could be retrieved.

Our model is bases on using concepts as indexing elements, enriching the external resource by new relations, and using the new relations with the old ones at matching time. The model consists of three main components:

3.1. *External resource*

The external resource is used in conceptual indexing to map a text to concepts. It contains: Terms $T = \{t_1, t_2, \ldots\}$, Concepts $C = \{c_1, c_2, \ldots\}$ and Relations $R = \{r | r \subseteq C \times C\}$ between concepts.

We enrich the external resource by:

1) Adding relations between concepts: theses added relations are computed automatically from the external resource. For example, "*shared-words*": this relation means that there are words in common between two concepts. We use this relation based on the hypothesis that, the more common words two concepts have, the more semantically close they are.

2) Defining a Certainty property to distinguish relations that are predefined in the external resource R_C from relations that are added $R_{\neg C}$. The Certainty represents how much it is sure that there is a semantic relation between two concepts. We distinguish these two types of relations $(R_C, R_{\neg C})$ departing from a hypothesis that, if there is a document d contains a concept c_d, a query q contains a concept c_q, and if there is a relation of type R_C (e.g. *isa*) between c_d and c_q. Then it is more probable that d is relevant document for q than if the relation between c_d and c_q is of type $R_{\neg C}$

(e.g. *shared-words*). $R = R_C \cup R_{\neg C}$ and $R_C \cap R_{\neg C} = \emptyset$. A new function is defined to calculate the certainty of a relation r:

$$\forall r \in R, \quad certainty\,(r) = \begin{cases} 1 & r \in R_C \\ x \in]0,1[& r \in R_{\neg C} \end{cases} \quad [8]$$

3) Defining the notion of '*Strength of relation*' which represents the ability of a relation to retrieve relevant documents of a query. The strength of relation is calculated by using the following formula:

$$\forall r \in R, \forall\,(c_i, c_j) \in r, \quad Strength_r\,(c_i, c_j) = sim_r\,(c_i, c_j) \times certainty\,(r) \quad [9]$$

Where:
$certainty\,(r)$ the certainty of a relation r.
$sim_r\,(c_i, c_j)$ the semantic similarity between two concepts.
Practically, there are many ways to calculate the semantic similarity between two concepts (Mohler *et al.,* 2009), but similarity calculation differ according to the relation between concepts.
By example, for "*isa*" relation, as it represents a hierarchical structure in an external resource, the similarity measure of *Leacock* can be used:

$$\forall\,(c_i, c_j) \in isa, \quad sim_{isa}\,(c_i, c_j) = -\log \frac{minLen\,(c_i, c_j)}{2 * L} \quad [10]$$

Where:
L is the depth of the isa-hierarchy of concepts
$minLen\,(c_i, c_j)$ is the isa-path of minimum length between c_i and c_j
And for "*shared-words*" relation, a variation of *mutual-information* measure can be used:

$$sim_{shared-words}\,(c_i, c_j) = \frac{\text{number of shared words between } c_i \text{ and } c_j}{\text{number of words in } c_i \times \text{number of words in } c_j} \quad [11]$$

Finally we define the conceptual indexing function $Index$: suppose there is a query q and a document $d \in D$ then:

$$Index : D \cup \{q\} \rightarrow C^* \quad [12]$$

where, C^* is the set of all subsets of C

3.2. *Bayesian Network*

To compute the matching between a document and a query, we use Bayesian network (Turtle *et al.,* 1990) (Turtle *et al.,* 1991) (Murphy, 1998) (Le, 2009). The network in our model contains three types of nodes: documents D, concepts C, and query q. Nodes are connected by using three types of weighted links (see Figure 1):

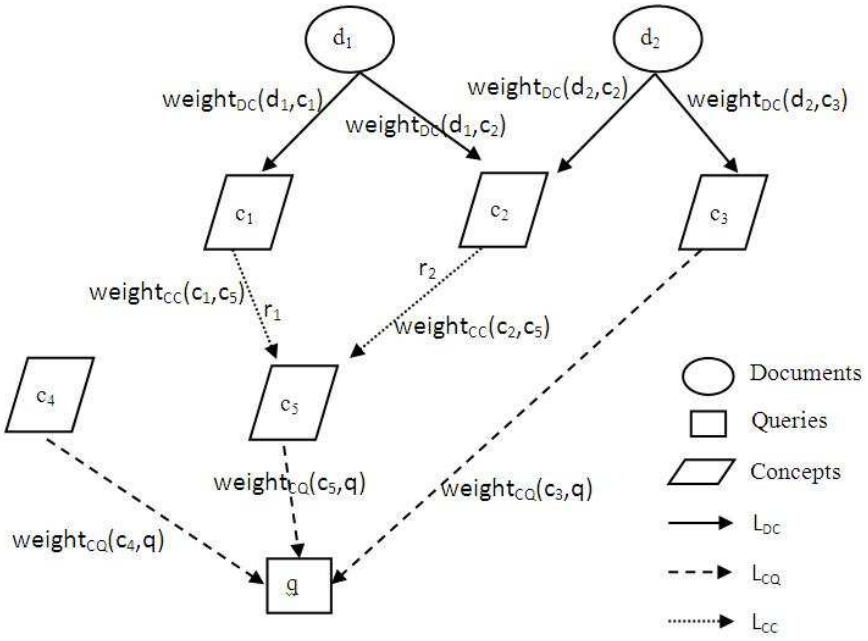

Figure 1. *Bayesian Network*

1)$L_{DC} = \{(d, c) \,|d \in D, c \in Index\,(d)\}$: links from documents to their concepts, weighted by $weight_{DC} : L_{DC} \rightarrow [0, 1]$ the importance of a concept in its document.
2) $L_{CQ} = \{(c, q) \,|c \in Index\,(q)\}$: links from concepts to their query, weighted by $weight_{CQ} : L_{CQ} \rightarrow [0, 1]$ the importance of a concept in the query.
3) $L_{CC} = \{(c_i, c_j) \,|\exists d \in D, c_i \in Index\,(d)\,, c_j \in Index\,(q)\,, \exists r \in R, (c_i, c_j) \in r\}$: links from documents' concepts to query's concepts, represent a relation between two concepts, weighted by $weight_{CC} : L_{CC} \rightarrow [0, 1]$ the strength of the relation between concepts.

Network construction: We construct Bayesian Network according to the following steps:
1) at indexing time:
a) For each document $d \in D$, we build a node d.
b) Indexing the document d.
c) For each concept $c \in Index\,(d)$, we build a node c and a link from d to c (if the node c is already built, we don't build another one, we just build a link).
2) at matching time:
a) For the query q, we build a node q.
b) Indexing the query q.

c) For each document $d_i \in D$:

i) For each concept $c_k \in Index(q)$, we build a node c_k and a link from c_k to q (if the node c_k is already built $c_k \in Index(d_i)$, we don't build another one, we just build a link).

ii) Building links from document's concepts $c_j \in Index(d_i) \wedge c_j \notin Index(q)$ to query's concepts $c_k \notin Index(d_i) \wedge c_k \in Index(q)$, if $\exists r \in R, (c_j, c_k) \in r$.

iii) we calculate RSV (Relevance Status Value) between the document $d_i \in D$ and the query q then we return to step $(2.c)$ to choose another document.

3.3. *Correspondence function*

To calculate RSV, we use the calculation rules of the conditional probability in Bayesian network according to the following steps:

1) choosing a document $d_{selected}$ from the documents collection D, then:

$$\forall d \in D, \quad P(d) = \begin{cases} 1 & d = d_{selected} \\ 0 & \text{else} \end{cases} \qquad [13]$$

2) for concepts that belong to the selected document $\{c_i | (d_{selected}, c_i) \in L_{DC}\}$:

$$P(c_i|L_{DC}) = \frac{weight_{DC}(d_{selected}, c_i) \times P(d_{selected})}{\sum_{(d_j,c_i)\in L_{DC}} weight_{DC}(d_j, c_i)} \qquad [14]$$

3) for concepts that belong to the query and don't belong to the selected document $\{c_i | c_i \in Index(q), c_i \notin Index(d_{selected}), \exists c_j \in Index(d_{selected}), (c_j, c_i) \in L_{CC}\}$:

$$P(c_i|L_{CC}) = \frac{\sum_{(c_j,c_i)\in L_{CC}} weight_{CC}(c_j, c_i) \times P(c_j|L_{DC})}{\sum_{(c_j,c_i)\in L_{CC}} weight_{CC}(c_j, c_i)} \qquad [15]$$

Note that if query's concept c_i doesn't have any relation with any document's concept c_j: $\forall c_j \in Index(d_{selected}), (c_j, c_i) \notin L_{CC}$ then $P(c_i|L_{CC}) = 0$.

4) now for the query node:

$$RSV(d_{selected}, q) = P(q|L_{CQ}) = \frac{\sum_{(c_i,q)\in L_{CQ}} weight_{CQ}(c_i, q) \times P(c_i|L_{CC})}{\sum_{(c_i,q)\in L_{CQ}} weight_{CQ}(c_i, q)} \qquad [16]$$

3.4. *Example*

To illustrate the calculation mechanism of RSV between a document d and a query q, we will present in this section a complete example. We will refer to concepts by their corresponding text for simplicity and clarity.

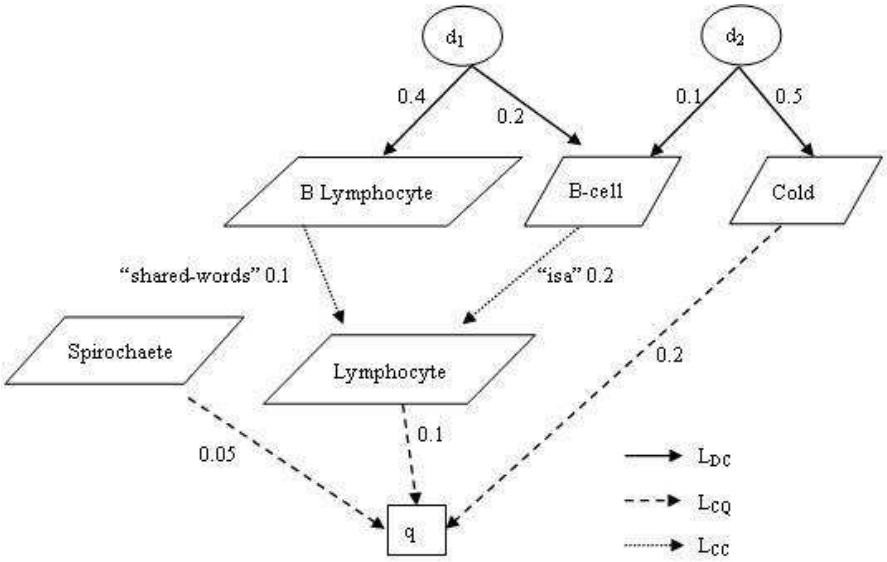

Figure 2. *Complete Example*

Suppose there is a Bayesian Network like (see Figure 2), and suppose the following weights:

$weight_{DC}\,(\text{"}BLymphocyte\text{"}, d_1) = 0.4$
$weight_{DC}\,(\text{"}B-cell\text{"}, d_1) = 0.2$
$weight_{DC}\,(\text{"}B-cell\text{"}, d_2) = 0.1$
$weight_{DC}\,(\text{"}Cold\text{"}, d_2) = 0.5$
$weight_{CQ}\,(\text{"}Lymphocyte\text{"}, q) = 0.1$
$weight_{CQ}\,(\text{"}Cold\text{"}, q) = 0.2$
$weight_{CQ}\,(\text{"}Spirochaete\text{"}, q) = 0.05$
$weight_{CC}\,(\text{"}B-cell\text{"}, \text{"}Lymphocyte\text{"}) =$
$\quad Strength_{isa}\,(\text{"}B-cell\text{"}, \text{"}Lymphocyte\text{"}) = 0.2$
$weight_{CC}\,(\text{"}BLymphocyte\text{"}, \text{"}Lymphocyte\text{"}) =$
$\quad Strength_{shared-words}\,(\text{"}BLymphocyte\text{"}, \text{"}Lymphocyte\text{"}) = 0.1$

For simplifying the formulas, we will use the following symbols: '*w*' instead of '*weight*', '*a*' instead of "*B Lymphocyte*", '*b*' instead of "*B-cell*", '*c*' instead of "*Cold*", '*d*' instead of "*Spirochaete*" and '*e*' instead of "*Lymphocyte*".

1) By choosing the document d_1, then:
$d_1 = d_{selected} \quad P(d_1) = 1 \quad P(d_2) = 0$
2) for concepts that belong to the selected document:
$P(b|L_{DC}) = \frac{w_{DC}(b,d_1)\times P(d_1)}{w_{DC}(b,d_1)+w_{DC}(b,d_2)} = \frac{0.2\times 1}{0.2+0.1} = 0.666$

$P(a|L_{DC}) = \frac{w_{DC}(a,d_1)\times P(d_1)}{w_{DC}(a,d_1)} = 1$
3) for concepts that belong to the query and doesn't belong to the selected document
$P(d|L_{CC}) = 0$ because $\forall c_j \in Index(d_{selected}), (c_j, d) \notin L_{CC}$
$P(e|L_{CC}) = \frac{w_{CC}(a,e)\times P(a|L_{DC})+w_{CC}(b,e)\times P(b|L_{DC})}{w_{CC}(a,e)+w_{CC}(b,e)} = \frac{0.1\times 1+0.2\times 0.666}{0.1+0.2} = 0.777$
$P(c|L_{CC}) = 0$ because $\forall c_j \in Index(d_{selected}), (c_j, c) \notin L_{CC}$
4) the correspondence value between the document d_1 and the query q:

$$\begin{aligned} RSV(d_1, q) &= P(q|L_{CQ}) \\ &= \frac{w_{CQ}(c,q)\times P(c|L_{CC})+w_{CQ}(e,q)\times P(e|L_{CC})+w_{CQ}(d,q)\times P(d|L_{CC})}{w_{CQ}(c,q)+w_{CQ}(e,q)+w_{CQ}(d,q)} \\ &= \frac{0.2\times 0+0.1\times 0.777+0.05\times 0}{0.2+0.1+0.05} = \frac{0.077}{0.35} = 0.22 \end{aligned}$$

3.5. *Conclusion*

We have presented in the previous subsections our model to solve term and concept mismatch problems. In this model, the documents and the query are represented by concepts, and we have also modeled the different relations between concepts.
This model depends also on the techniques of Bayesian Network to compute the correspondence value between a document and a query.

4. Model validation

4.1. *Validation context*

The proposed model is validated by applying it to the test collection: ImageCLEFMed2005, and by using the UMLS 2005 as an external resource. We use MetaMap (Aronson, 2006) tool to identify concepts from raw text, we program a tool to build Bayesian network and calculate correspondence value, and we use the *tf.idf* measure to calculate the importance of a concept in its document. We use this measure because it is useful to compute how much a concept describes a document (the part *tf* of the measure) and at the same time, how much a concept discriminates a document in a collection (the part *idf* of the measure).

ImageCLEFMed is a part of CLEF (Cross-Language Evaluation Forum), which is a yearly campaign for evaluation of multilingual information retrieval since 2000. ImageCLEFMed concerns searching medical images depending on heterogeneous and multilingual documents that contain text and images. The test collection ImageCLEFMed 2005 contains *Casimage (9000 images), MIR (2000 images), PEIR (33000 images),* and *PathoPic (9000 images)* (Clough *et al.*, 2005). This collection includes more than *50000* images and their annotations in XML format. Majority of these annotations are English text, but there are French and German text also. The collection contains 25 queries, and each query is written in the three languages (English, French, German).

UMLS is a multi-source knowledge base in the medical domain. It contains three sources of knowledge:
1- Metathesaurus: is a vocabulary database in the medical domain, extracted from many sources, each source of them is called "Source Vocabularies". The Metathesaurus is organized in Concepts, which represent the common meaning of a set of strings extracted from different source vocabularies.
2- Semantic Network: consists of a set of Semantic Types linked together by two different types of Semantic Relations (hierarchical, non- hierarchical). The purpose of the Semantic Network is to provide a consistent categorization of all concepts represented in the UMLS Metathesaurus.
3- SPECIALIST Lexicon: is a set of general English or biomedical terms and words extracted from different sources.
Moreover, UMLS contains many tools to deal with these different sources (e.g. MetamorphSys, UMLS Knowledge Source Server).

MetaMap is a tool to map text to UMLS concepts. This tool is composed of the following components:
1- Morphology and Syntax: extraction of noun phrases from text using NLP techniques.
2- Variation: construction of different forms (variants) of the noun phrase or part of it.
3- Identification: for each noun phrase variant, it retrieves all concepts that possibly correspond to this variant. The set of concepts that possibly corresponds to the noun phrase, is called "Candidates set".
4- Evaluation: ordering the concepts of candidate set according to an evaluation function (f), which determines: "how much the concept represents the noun phrase?"
5- Disambiguation: reduction of the size of the candidates set.

We want to show if the enrichement of the external resource helps to retrieve more relevant documents.

4.2. *Model variants*

We have tested the following variants of the model:

1) BASE: there is no relations between concepts, i.e. the matching between a document and a query depends on the shared concepts between them. BASE is an example of Intersection-based IRSs.
2) 2REL-C, 2REL-L, 2REL-E: two relations (*isa*, *part-of*: these relations are predefined in UMLS) are used to link documents' concepts and query's concepts.
From one side,
$isa \in R_C, part-of \in R_C$ then $certainty\,(isa) = certainty\,(part-of) = 1$
From another side, these three variants use different similarity functions,
- in 2REL-C: $sim_{isa} = sim_{part-of} = \alpha$ where α is a constant. After many

experiments, we found that the value of α that gives the best results is $\alpha = 0.2$.
- in 2REL-L: a normalized version of *Leacock* measure is used

$$sim_{isa}(c_i, c_j) = sim_{part-of}(c_i, c_j) = \frac{-\log \frac{minLen(c_i,c_j)}{2*L}}{-\log \frac{1}{2*L}} \quad [17]$$

- in 2REL-E: a version of the exponential function is used

$$sim_{isa}(c_i, c_j) = sim_{part-of}(c_i, c_j) = e^{-minLen(c_i,c_j)} \quad [18]$$

3) 3REL-C, 3REL-L, 3REL-E: beside the two relations (*isa*, *part-of*) that are used in the 2REL variants, a new relation (*shared-words*: this relation is added to UMLS) is used. Here,
$shared - words \in R_{\neg C}$ then $certainty(shared - words) \in]0, 1[$ and after many experiments, we found that the value that gives the best results is $certainty(shared - words) = 0.1\ (10\%)$
In addition, in 3REL variants, the similarity of *isa* and *part-of* relations is estimated by using the same functions as in 2REL variants, and the similarity of *shared-words* relation is estimated by using a variant of mutual-information measure

$$sim_{shared-words}(c_i, c_j) = mutual - information(c_i, c_j) \quad [19]$$

4.3. *Results and Discussion*

We got the following results:

Table 2. *MAP, Precision at first 5 and R-Precision of BASE, 2REL, 3REL*

	MAP	P@5	R-Prec
BASE	0.1240	0.1920	0.1634
2REL-C	0.1327	0.1920	0.1698
2REL-L	0.1349	0.1840	0.1636
2REL-E	0.1352	**0.2000**	**0.1777**
3REL-C	0.1358	0.1920	0.1691
3REL-L	0.1363	0.1760	0.1620
3REL-E	**0.1384**	**0.2000**	0.1774

From the previous results and by comparing the following variants (BASE, 2REL-x, 3REL-x) [4]:

First, more relevant documents for a query could be retrieved by using relations between concepts (see Table 3). In addition, the enrichment of the external resource

4. x means C, L or E

Table 3. *Number of relevant, retrieved, and retrieved-relevant documents of BASE, 2REL, 3REL*

	# Relevant documents	# Retrieved documents	# Retrieved-Relevent
BASE	2217	58037	1234
2REL-x	2217	119896	1474
3REL-x	2217	144789	**1698**

by a very simple relation "*shared-words*" and using it, allows the system to retrieve even more relevant documents (see Table 3).

Second, the average precision of the system (MAP) is increased by using relations (see Table 2).

$$\begin{array}{l} MAP_{3REL-C} > MAP_{2REL-C} > MAP_{BASE} \\ MAP_{3REL-L} > MAP_{2REL-L} > MAP_{BASE} \\ MAP_{3REL-E} > MAP_{2REL-E} > MAP_{BASE} \end{array} \quad [20]$$

The problem in the two variants (2REL-C, 3REL-C) is the use of a constant value α to estimate the similarity between two concepts. Using constant value has two main drawbacks: **1)** learning phase is needed to estimate the value of α, **2)** the similarity is constant whatever the length of the path between the two concepts. Normally, the similarity should be decreased when the length of the path is increased.

To avoid the problems of the constant measure, a normalized version of *Leacock* measure is used in the two variants (2REL-L, 3REL-L). By using normalized *Leacock* measure, the average precision of the system is increased (see Table 2).

$$\begin{array}{l} MAP_{2REL-L} > MAP_{2REL-C} \\ MAP_{3REL-L} > MAP_{3REL-C} \end{array} \quad [21]$$

However, the precision at the first 5 documents (P@5) and the exact precision (R-Prec) are decreased (see Table 2).

$$\begin{array}{l} P@5_{2REL-L} < P@5_{2REL-C} \\ P@5_{3REL-L} < P@5_{3REL-C} \end{array} \quad [22]$$

$$\begin{array}{l} R-Prec_{2REL-L} < R-Prec_{2REL-C} \\ R-Prec_{3REL-L} < R-Prec_{3REL-C} \end{array} \quad [23]$$

We think that happened because the normalized *Leacock* measure has two main drawbacks: **1)** if $c_i = c_j$ then $minLen\,(c_i, cj) = 0$, in this case $sim_{isa}\,(c_i, cj)$ and $sim_{part-of}\,(c_i, cj) \rightarrow +\infty$ and this is not practical, **2)** if $c_i \xrightarrow{isa} c_j$ or $c_i \xrightarrow{part-of} c_j$ then $minLen\,(c_i, c_j) = 1$, in this case $sim_{isa}\,(c_i, cj) = sim_{part-of}\,(c_i, cj) = 1$, but even with a direct relation (a relation of length one), there is a penalty, so the similarity should be less than 1.

To avoid the problems of normalized *Leacock* measure, the exponential function is used in the two variants (2REL-E, 3REL-E).

First, by using the exponential function, if $minLen\,(c_i, c_j) = 0$ then $sim_{isa}\,(c_i, cj) = sim_{part-of}\,(c_i, cj) = 1$, and if $minLen\,(c_i, c_j) = 1$ then $sim_{isa}\,(c_i, cj)$ and $sim_{part-of}\,(c_i, cj) < 1$.

Second, the average precision of the system, the precision at the first 5 documents and the exact precision all are increased (see Table 2).

$$\begin{array}{c} MAP_{2REL-E} > MAP_{2REL-L} > MAP_{2REL-C} \\ MAP_{3REL-L} > MAP_{3REL-L} > MAP_{3REL-C} \end{array} \quad [24]$$

$$\begin{array}{c} P@5_{2REL-E} > P@5_{2REL-C} > P@5_{2REL-L} \\ P@5_{3REL-E} > P@5_{3REL-C} > P@5_{3REL-L} \end{array} \quad [25]$$

$$\begin{array}{c} R-Prec_{2REL-E} > R-Prec_{2REL-C} > R-Prec_{2REL-L} \\ R-Prec_{3REL-E} > R-Prec_{3REL-C} > R-Prec_{3REL-L} \end{array} \quad [26]$$

Finally, by using relations, more relevant and irrelevant documents could be retrieved (see Table 3), but in spite of that, the precision of the system is enhanced. That's mean, the relevant documents ranked well comparing to the irrelevant documents.

From the previous results, using relations contributes in solving concept mismatch problem, and extending the external resource by new relations, even a very simple relation, could contribute more in solving the problem.

5. Conclusion

We showed in this work that conceptual indexing is insufficient to solve the term mismatch problem. The use of relations from the conceptual resource increase the MAP, but we think that in UMLS, too many potential relations between concepts are missing. When we add these relations, we showed an interesting increase in the MAP. In conclusion, this research tend to show that existing resources even very larges ones like UMLS, are not totally adapted to IR because of the lack of relations between concepts. This lack can be compensated by analysis of terms associated to concepts.

Finally there are many points in this work, that need more study, like studying the influence of adding other relations to the model, using properties other than Certainty to describe relations, and validation the model by using another test collections and another external resources.

6. References

Aronson A. R., « Metamap: Mapping text to the umls metathesaurus », 2006.

Baziz M., Indexation conceptuelle guidée par ontologie pour la recherche d'information, Thèse de doctorat, Université Paul Sabatier, Toulouse, France, décembre, 2005.

Bodenreider O., Burgun A., Botti G., Fieschi M., Le Beux P., Kohler F., « Evaluation of the Unified Medical Language System as a medical knowledge source. », *J Am Med Inform Assoc*, vol. 5, n° 1, p. 76-87, 1998.

Bodenreider O., Burgun A., Rindflesch T. C., « Lexically-suggested hyponymic relations among medical terms and their representation », *in the UMLS, in Proceedings of TIA2001, 1121*, 2001.

Chevallet J.-P., « endogènes et exogènes pour une indexation conceptuelle intermédia », Mémoire d'Habilitation a Diriger des Recherches, 2009.

Chevallet J.-P., Lim J. H., Le T. H. D., « Domain Knowledge Conceptual Inter-Media Indexing, Application to Multilingual Multimedia Medical Reports », *ACM Sixteenth Conference on Information and Knowledge Management (CIKM 2007), Lisboa, Portugal*, November 6–9, 2007.

Clough P., Müller H., Deselaers T., Grubinger M., Lehmann T. M., Jensen J. R., Hersh W. R., « The CLEF 2005 Cross-Language Image Retrieval Track », *CLEF*, p. 535-557, 2005.

Crestani F., « Exploiting the similarity of non-matching terms at retrieval time », *Journal of Information Retrieval*, vol. 2, p. 25-45, 2000.

Crestani F., Rijsbergen C. J. V., « Information retrieval by logical imaging », *Journal of Documentation*, vol. 51, p. 3-17, 1995.

Le T. H. D., Utilisation de ressource externes dans un modèle Bayesien de Recherche d'Information: Application a la recherche d'information médicale multilingue avec UMLS, PhD thesis, Université Joseph Fourier, Ecole Doctorale MSTII, 2009.

Maisonnasse L., Les supports de vocabulaires pour les systèmes de recherche d'information orientés précision : application aux graphes pour la recherche d'information médicale., PhD thesis, Université Joseph-Fourier - Grenoble I, 05, 2008.

Mohler M., Mihalcea R., « Text-to-Text Semantic Similarity for Automatic Short Answer Grading », *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, Association for Computational Linguistics, Athens, Greece, p. 567-575, March, 2009.

Murphy K., « A Brief Introduction to Graphical Models and Bayesian Networks », 1998.

Turtle H., Croft W. B., « Inference Networks for Document Retrieval », p. 1-24, 1990.

Turtle H., Croft W. B., « Evaluation of an Inference Network-Based Retrieval Model », *ACM Transactions on Information Systems*, vol. 9, p. 187-222, 1991.

Woods W., « Conceptual Indexing: A Better Way to Organize Knowledge », 1997.