Query expansion using medical information extraction for improving information retrieval in French medical domain

Aicha GHOULAM*, Fatiha BARIGOU*, Ghalem BELALEM*, Farid MEZIANE**
**Department of Computer Science,
University of Oran 1, Ahmed Ben Bella, Oran, Algeria
**Informatics Research Centre, Newton Building,
University of Salford, M5 4WT, UK*

ABSTRACT

Many users' queries contain references to named entities, and this is particularly true in the medical field. Doctors express their information needs using medical entities as they are elements rich with information that helps to better target the relevant documents. At the same time, many resources have been recognized as a large container of medical entities and relationships between them such as clinical reports; which are medical texts written by doctors. In this paper, we present a query expansion method that uses medical entities and their semantic relations in the query context based on an external resource in OWL. The goal of this method is to evaluate the effectiveness of an information retrieval system to support doctors in accessing easily relevant information. Experiments on a collection of real clinical reports show that our approach reveals interesting improvements in precision, recall and MAP in medical information retrieval.

Keywords: Information Retrieval, Query Expansion, OWL external resource, Medical Entities Recognition, Semantic Relations Extraction.

INTRODUCTION AND MOTIVATION

With the growing amount of available information in the medical field, accessing useful and relevant biomedical information in real time is becoming of a paramount importance for practitioners and researchers. Indeed, information retrieval systems support users in their daily activities to satisfy their needs. Usually, the user formulates his information need into a query; in return, an information retrieval system (IRS) provides the most relevant documents that satisfy the user query. However, there are many difficulties in developing an effective IRS. One of these difficulties is the word mismatching problem (vocabulary mismatch). The users can express their needs using different words with similar meanings (synonyms) and same words with different meanings (polysemy). According to (Bhatnagar & Pareek, 2014), concepts may be described in different words in user's queries and/or documents. Many techniques were proposed to solve this problem; one of them is query expansion techniques.

For a long time, query expansion (QE) has been the main motivation for improving the retrieval effectiveness of an IRS. The QE can be performed in different ways such as manual (the user chooses expansion terms), interactive (the user chooses expansion terms from suggestions provided by the system) and automatic (all the process is invisible to the user).

Researchers developed efficient techniques for automatic query expansion; a survey of these techniques is given in (Carpineto & Romano, 2012). Sources for selecting the query expansion terms can be grouped into: document corpus (global, local, relevance feedback-based query expansion), linguistic resources (dictionaries, thesaurus, WorldNet ontology, for semantic query expansion) and world knowledge-based resources (Wikipedia). Recently, systems based on query expansion make use of external resources such as ontologies and lexical hierarchies and they have significantly improved their results. In the medical field, most of the medical ontologies such as Medical Subject Heading (MeSH) thesaurus were used to improve medical information retrieval. In (Díaz-Galiano et al., 2009) terms associated with MeSH descriptors are considered as synonyms and used to expand queries, the experiments have shown improvements in the performance of the information retrieval in the medical domain.

In medicine, most of the ontologies have been translated to French to cover the general concepts in the domain. For example, the Health Terminology/Ontology Portal (HeTOP[1]) contains different medical ontologies translated to French such as Medical Subject Headings (MeSH), Systematized Nomenclature of MEDicine (SNOMED int), National Cancer Institute Thesaurus (NCIt) and so on. However, these ontologies are too large to be processed in a specific system. Thus, a domain-specific ontology is needed to solve this problem. This is what led us to construct our own specific resource in OWL and then use it to expand user's query.

In this paper, we used an external resource in OWL for query expansion process in the French medical domain. It contains medical entities and relations between them that were extracted from real clinical reports and improved from the work developed in (Ghoulam et al., 2015b). We used medical entities, their synonyms and the semantic relations between them to expand the user's query. Then, we transformed the expanded query to Boolean query using Boolean operators.

This work is motivated by the fact that, the clinical reports can have a positive impact on the quality of care, patient safety and efficiencies in medical procedures. The doctors need a quality search, to consult and search through these informative reports so that they can make a decision in the shortest time to improve healing. These kinds of medical retrieval systems have become very necessary tools; they will enable researchers to access accurate data and the required information and reducing the time spent by doctors on making decisions about patient's diseases.

The remainder of this paper is organized as follows; section 2 presents related works on information retrieval systems and query expansion methods in the medical field. In section 3, we describe the proposed system and the external resources we used. Our contribution to retrieve medical reports using query expansion will be discussed in section 4. Section 5 presents the experiments and their results. The paper ends with a conclusion and some suggestions on future developments.

RELATED WORKS

In general, named entities are elements rich with information, hence, several researches are interested to classify them (Nadeau & Sekine, 2007) or to disambiguate them (Hoffart et al., 2011). In information retrieval field, named entities were used for indexing (Buizza, 2011) and for query expansion (Audeh et al., 2014). In medical field, many studies were developed to recognize medical entities (Ghoulam et al., 2015a), (Barigou et al., 2012) and to extract the semantic relations between them (Ben Abacha, 2011).

At the present time, semantic information retrieval becomes an important part of any information processing engine. The semantic annotations are usually described by ontology, which is an explicit specification of a conceptualization of things. It plays a vital role in describing the semantic information. Several works aim to construct an ontology, (Suresh & Zayara, 2014) used a syntactic and semantic probability-based naive Bayes classifier to extract concept relations from the unstructured text for the automatic construction of ontology, for that a list of attributes and associations of the given seeds concept are automatically extracted. In reference (Bentricia et al., 2017) proposed an approach based on conjunctive patterns to extract semantic relations from Quranic Arabic corpus to enrich automatic construction of Quran ontology. The role to construct such ontology is to provide precise and comprehensive knowledge to the world, with the aim to reduce the role of expert knowledge to built ontology. (Denis & Wasito, 2017) proposed a fully automatic method that combines two approaches (ontology learning from texts and ontology design pattern) to built bilingual domain ontology precisely to Alzheimer's domain knowledge.

To allow ontologies to be machine processable, their modeling is often implemented in a different language such as Resource Description Framework (RDF) or Web Ontology Language (OWL). To achieve the disease intelligence through the web, (Prabath & Saluka, 2012) proposed a methodology based on ontologies created using OWL. They created ontology named disease ontology by extracting information about rapid spreading and changing diseases, they used OWL to represent the knowledge such as concepts, terms and relationships. Samwald et al. (2013) proposed a semantic knowledge-base relying on RDF and OWL technologies, in order to manage data for clinical pharmacogenetics, An RDF model is used to capture detailed results of manual annotation; the OWL ontology contains the detail of drug labels of pharmacogenomics information.

Generally, ontologies are used in several applications such as query expansion. The query expansion is used in an IRS when new terms are added to the user's query in order to improve and increase the effectiveness of the retrieval process. Researchers explore query categorization for query expansion; the taxonomy of query classes was given in (Dipasree et al., 2015). Related works on query expansion can be broadly classified into three groups: global, local and external. In global query expansion, the entire corpus is considered for selecting the expansion terms. A global technique was proposed in (Jing & Croft, 1994) based on term co-occurrence information in the corpus, they select expansion terms that are most similar to the query. In local QE techniques, the terms are selected from an initial set of documents retrieved in response to the original query for example relevance feedback (Bilel et al., 2011; Picariello et al., 2007). In the absence of user feedback, a few top-ranked documents are assumed to be relevant this is called pseudo-relevance feedback. Pragati et al. (2014) improved the limitation of pseudo-relevance feedback query expansion by suggesting a hybridization of corpus-based information with a genetic fuzzy approach and semantic similarity notion. Colace et al. (2015) proposed a new method to expand query based on weighted word pairs approach. This structure is extracted from the set of documents obtained through the relevance feedback and then added to the initial query. In external QE techniques, researchers incorporate the notion of semantics by the use of external linguistic knowledge like WordNet ontology. Abbache et al. (2014) used Arabic WordNet for Arabic query expansion by adding the synonyms of terms to the original query; the method doesn't give good results comparing to the interactive method therefore in (Abbache et al., 2016) they used method that select automatically synonyms extracted from Arabic WordNet based on association rules, this method improves the results of retrieval regarding the mean average precision (MAP). The experiments show that with a good method of synonyms selection, the use of Arabic WordNet as a source of linguistic information for automatic query expansion improves the effectiveness of Arabic information retrieval. A general ontology was used in (Audeh et al., 2014) for named entities expansion using the "Yago" ontology.

Many researchers are being active in the medical domain. Hersh & Hickam (1995) identified an optimal approach to index, retrieve and evaluate resources in the biomedical domain under the name of the Saphire project. Mohameth et al. (2012) proposed a method based on relevance feedback and MeSH ontology for the query expansion. Chen et al. (2016) proposed an approach to semantic expansion system based on medical ontology (Hepatitis ontology). They construct Hepatitis ontology for querying. They focus on three semantic expansion query including synonym expansion, hypernym/hyponym expansion and expansion of similar words. For medical question answering system, Embarek & Ferret (2012) extracted medical entities and relations between them to construct a medical ontology. The ontology was used for a question answering system to respond to medical questions. Similarly, Ben Abacha & Zweigenbaum (2015) used semantic approach for medical question answering. They extract medical information that represented answers to their questions.

A comparison between our work and the work done by (Ben Abacha et al., 2015) is shown in Table 1 below. It differs on the following points: the corpus's language, the methods of medical entities recognition, the methods of semantic relations extraction, and the language use in building knowledge-base, the system building, the corpus used, the external knowledge-base and the purpose of the constructed system.

	System MEANS.	Our work
Language	English	French
Medical entity recognition methods	MetaMap Plus, SVM, Bio-CRF	Local Grammar
Semantic relation extraction methods	Pattern-based approach, SVM	Pattern-based approach
Language use in building knowledge base	RDF	OWL
Designed system	Question Answering system	Medical information retrieval system: Query expansion
Data set (corpus used)	Medline, PubMed Challenge i2b2 2010	Real clinical reports
External knowledge base	DrugBank, Bio2RDF, BioGateWay	HeTop: SNOMED int., MeSH, NCIt…
Purpose of the construct system	Precise Answers to Medical Questions	Set of clinical reports to a medical query

Table 1. Comparing systems.

Multiple standardized ontologies are available in the medical field and were used in information extraction and retrieval tasks (e.g. UMLS, SNOMED CT, and MeSH). Jovic et al. (2007) proposed to represent medical knowledge through ontologies and provided a detailed process. Gangemi et al. (1998) classify ontologies according to the level of explicitness and formalization. Nowadays, researchers are interested in constructing a medical ontology for the semantic representation of knowledge. Charlet et al. (2012) developed medical ontology based on clinical reports and specialized thesaurus reuse such as CIM-10, CCAM, and SNOMED V3.5. To avoid redundancy in building a new ontology from scratch, Zulkarnaine et al. (2016) proposed a methodology to develop a new medical ontology by reusing existing biomedical ones such as FMA, SNOMED-CT, and RadLex.

For medical information retrieval systems, (Khadim et al., 2014) proposed to use external resource for improving information retrieval in the biomedical domain. They expand user's query by the use of controlled vocabularies such MeSH and UMLS. They used web pages from medical websites for the evaluation of the system. The experiments show that the query expansion methods outperform the baseline. Also (Michael et al., 2016) design and implement an efficient e-healthcare information retrieval system. They created ontology for human disease-treatment relationships and used WordNet to obtain the related terms which are semantically associated with the e-healthcare domain. Yangyang et al. (2017) proposed a semantic-based multi-analysis approach for medical information retrieval. This approach based on medical ontology MeSH, experiments on PubMed medical article collections show that this approach is feasible and efficient compared to other traditional approaches in medical retrieval.

They used biomedical literature on “hypertension” from PubMed. They didn't expand and exploit the semantic relations in their approach.

Our own work differs from the related work in the following aspects. First, we use French real clinical reports as a corpus in medical information retrieval system. Second, we construct our own medical ontology in OWL based on information extraction (Ghoulam et al., 2015a). Third, we used existing medical ontologies to enrich our medical ontology. Forth, we expand and exploit the semantic relations in our approach. On the other hand, we make three contributions in our work. First, we extract information from real clinical reports such as medical entities (Ghoulam et al., 2015b) and semantic relations between them. Second, we save this information as OWL annotations, and enrich it using different websites, for instance, Health Terminology/Ontology Portal (HeTop). Third, we use the OWL annotations in information retrieval system to expand the user’s query in medical domain. For the query expansion method; we used medical entities and semantic relations according to their nature and their context in the query to reformulate the expanded query as a Boolean query. An evaluation of French real clinical reports shows that the use of information extraction improves the performances in medical information retrieval. In the following section, we will describe our system architecture and the external resource used.

INFORMATION RETRIEVAL SYSTEM ARCHITECTURE

Before showing our query expansion approach; we will present our information retrieval system architecture as shown in Figure 1. As any IRS, our system has many phases: the indexation phase, the search phase, the analysis phase and the query expansion phase.

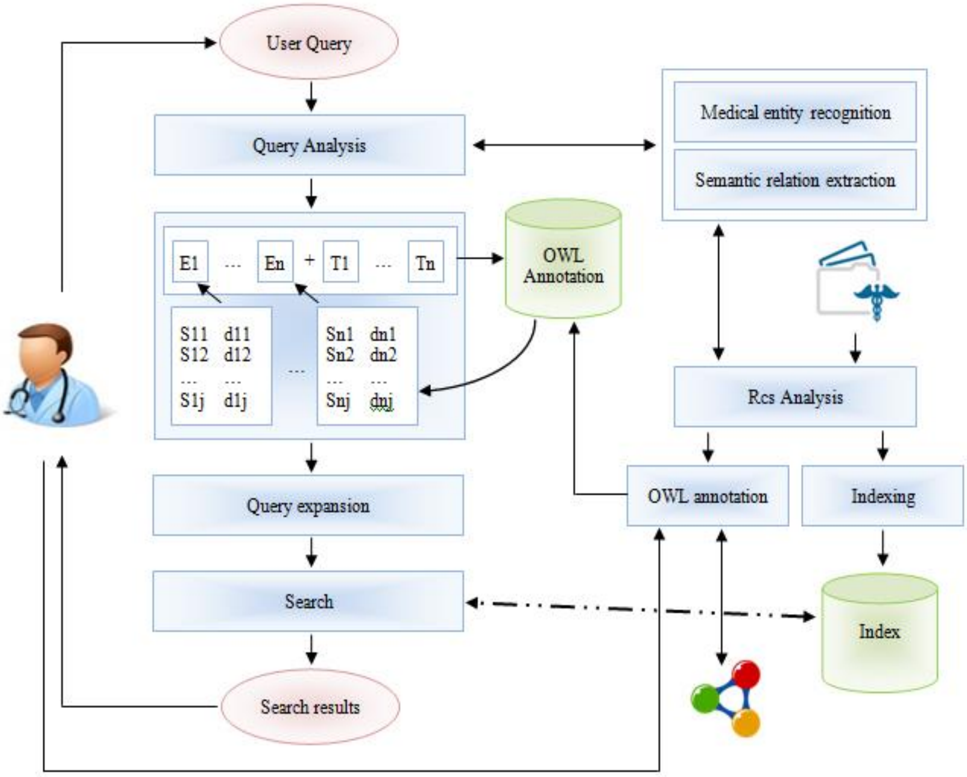


Figure 1. Architecture of proposal system.

The indexation and search phases

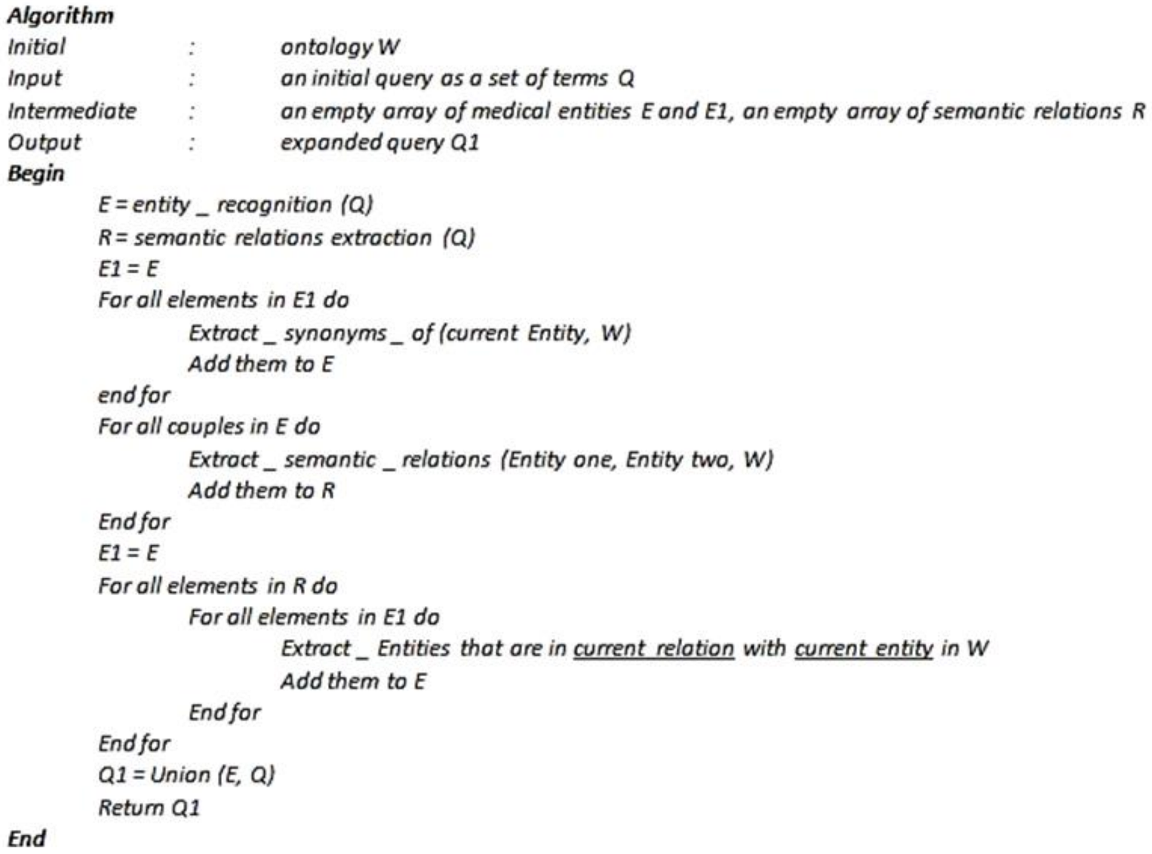
We used an open source library named Lucene[2] in the indexing and searching phases. To properly index the clinical reports, we used a French Analyzer that pre-process the texts data and convert them into terms stored in the index. The process can be summarized as follows: the indexing phase, which is responsible for indexing the clinical reports to build an index, and the searching phase, which is responsible for searching this index to retrieve the relevant reports.

The Analysis and query expansion phases

In the query analysis phase, we extract medical entities from the users' query. The same method as in (Ghoulam et al., 2015b) has been used for the extraction. The output of the analysis phase is a set of medical entities (E_1, E_2 ... E_n), a set of semantic relations (R_1, R_2 ... R_m) and a set of terms (T_1, T_2 ... T_k) that were not recognized as medical entities or semantic relations.

In the query expansion phase, we rely on the external resource (OWL annotation) to expand each medical entity E_i (i = 1 ... n) with their synonyms (S_{i1}, S_{i2} ... S_{ij}) and their descendants (hyponyms) (d_{i1}, d_{i2} … d_{il}), we add them all to the initial set of medical entities E. And for each couple in the set of medical entities (E_x,E_y) we extract the semantic relations between them using the same external resource, we add them to the initial set of semantic relations R. Now for each relation R_i (i = 1 ... m) in R and each medical entity E_j (j = 1 ... n+f) in E we extract all entities that have relation R_i with entity E_j. We add them to E. The expanded query is the union of E and T, without taking in consideration the duplication.

The algorithm of the query expansion process is shown in Figure 2:



```
Algorithm
Initial         :    ontology W
Input           :    an initial query as a set of terms Q
Intermediate    :    an empty array of medical entities E and E1, an empty array of semantic relations R
Output          :    expanded query Q1
Begin
    E = entity _ recognition (Q)
    R = semantic relations extraction (Q)
    E1 = E
    For all elements in E1 do
        Extract _ synonyms _ of (current Entity, W)
        Add them to E
    end for
    For all couples in E do
        Extract _ semantic _ relations (Entity one, Entity two, W)
        Add them to R
    End for
    E1 = E
    For all elements in R do
        For all elements in E1 do
            Extract _ Entities that are in current relation with current entity in W
            Add them to E
        End for
    End for
    Q1 = Union (E, Q)
    Return Q1
End
```

Figure 2. Algorithm of expansion.

The external resource (OWL annotation) that we used contains a set of medical entities and relations between them extracted from real clinical reports. Because of the lack of clinical reports and because it does not cover the entire field, we enriched this resource using a set of websites and validated by expert doctors.

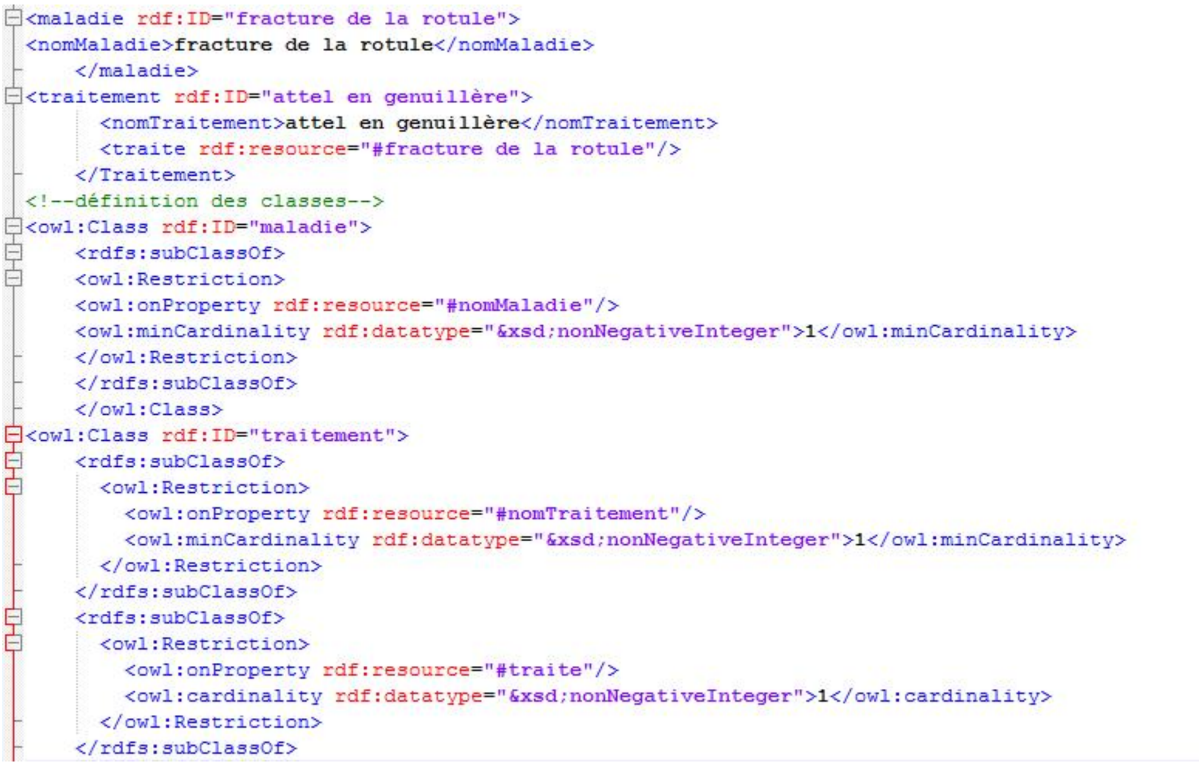
Table 2 summarizes the content of this resource, Table 3 gives a general structure of it and Figure 3 displays a portion of the resource in OWL.

	OWL annotation
Language	French
Number of medical entities	140
Number of semantic relations	65
Name of medical entities types	maladie, traitement, symptom, examen, medicament
Name of relation types	traite, detecte, soign, signe
Extraction type	Multiple word extraction and concept

Table 2: external resource (OWL annotation) representation.

```
<TypeOftargetEntiy  rdf:ID=”IDOf targetEntity”>
      <nomtargetEntity> … <nomtargetEntity>
</TypeOftargetEntiy >
<TypeOfsourceEntity  rdf:ID= ” IDOf sourceEntity >
     <nomsourceEntity> … <nomsourceEntity>
    <relationName  rdf:resource=”# nomtargetEntity”/>
</ TypeOfsourceEntity >
```

Table 3: overall structure of the OWL annotation.



```
<maladie rdf:ID="fracture de la rotule">
<nomMaladie>fracture de la rotule</nomMaladie>
    </maladie>
<traitement rdf:ID="attel en genuillère">
      <nomTraitement>attel en genuillère</nomTraitement>
      <traite rdf:resource="#fracture de la rotule"/>
    </Traitement>
<!--définition des classes-->
<owl:Class rdf:ID="maladie">
    <rdfs:subClassOf>
    <owl:Restriction>
    <owl:onProperty rdf:resource="#nomMaladie"/>
    <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:minCardinality>
    </owl:Restriction>
    </rdfs:subClassOf>
    </owl:Class>
<owl:Class rdf:ID="traitement">
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty rdf:resource="#nomTraitement"/>
        <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:minCardinality>
      </owl:Restriction>
    </rdfs:subClassOf>
    <rdfs:subClassOf>
      <owl:Restriction>
        <owl:onProperty rdf:resource="#traite"/>
        <owl:cardinality rdf:datatype="&xsd;nonNegativeInteger">1</owl:cardinality>
      </owl:Restriction>
    </rdfs:subClassOf>
```

Figure 3. Part of the OWL annotation.

QUERY EXPANSION METHODS

In this section, we present our approach for query expansion based on medical external resource. In fact, we apply three query expansion methods in order to reformulate the initial query. In the first method, we expand the medical entities found in the initial query with all their synonyms and descendants. In the second one, we expand the medical entities found in the initial query with all their synonyms, descendants and the semantic relations in the context of the query. Finally, in the third approach, we combined the first two methods into a Boolean query reformulation.

Medical entity expansion

In this method, we use medical entities found in the initial query for query expansion based on the external resource as shown in Figure 4.

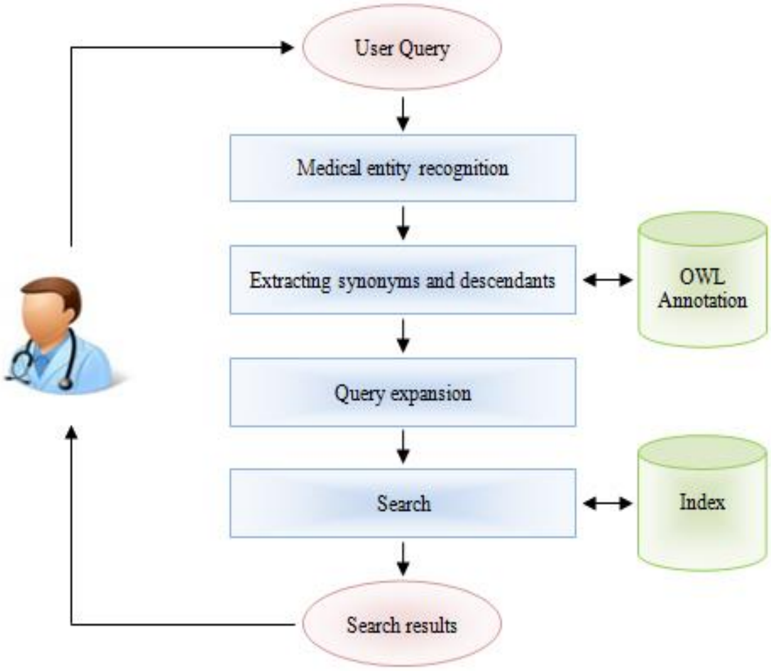


Figure 4. The medical entity expansion method.

This method is composed of four steps: medical entity recognition, extracting synonyms, query expansion, and index searching. Each of these steps is discussed in the following:

- **Medical entity recognition:** Doctors express their information needs using medical terms and entities and that is why it is important to recognize them. In our previous work (Ghoulam et al., 2015b) we used a rule-based approach to medical entity recognition. In this step, we used the same approach for the medical entity recognition.
- **Synonyms and descendants extraction:** the output of the previous step is a set of medical entities. In this step, our system find the synonyms and their descendants (hyponyms) for each medical entity recognized from the external resource which contains medical entities extracted from real clinical reports (Ghoulam et al., 2015b) and their relationships. And it was validated by medical doctors after enriching it using websites.
- **Query expansion:** medical entities used by a user in a query are not always sufficient to describe his needs. He will only get documents that contain the medical entities that are present in his query. We used the external resource to expand the original query using two semantic relations, namely synonym, and hyponym. For example, if a user writes the query "*traumatisme lombaire*", ("lumbar trauma") which is the name of a disease, it may be expanded to include "*traumatisme lombaire*", "*rachi lombaire*", "*fracture de L01*", "*fracture de L02*"; ("lumbar trauma", "lumbar spinal", "fracture L01", "fracture L02").
- **Searching in the index:** with the reformulation of the query, the synonyms and hyponym will be identified by the system that will return the reports containing those new entities.

Semantic relation extraction

This method has the same steps as the medical entity expansion method. The main difference is an extra step that concerns the extraction of semantic relations in the query context, as shown in Figure 5. Hence, the query reformulation is based on medical entity expansion and relation extraction.

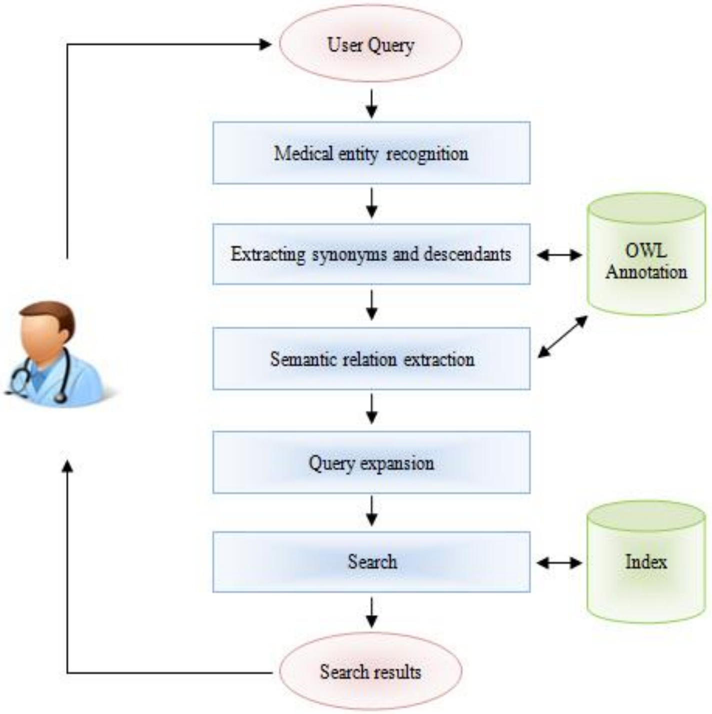

User Query
Medical entity recognition
Extracting synonyms and descendants
OWL Annotation
Semantic relation extraction
Query expansion
Search
Index
Search results

Figure 5. The semantic relation extraction method.

For example: if a user enters the query "*traitement de traumatisme du rachi lombaire*", ("treatment of Lumbar spinal trauma");

- The query passes through the medical entity recognition phase to recognize "*traumatisme du rachi lombaire*"; ("Lumbar spinal trauma") as a disease.
- The synonyms and descendants of the disease are extracted from the external resource (OWL Annotation), so the query will be expanded to include: "*traumatisme du rachi lombaire*", "*traumatisme lombaire*", "*rachi lombaire*", "*fracture de L01*", "*fracture de L02*"; ("Lumbar spinal trauma", "lumbar trauma", "lumbar spinal", "fracture L01", "fracture L02"). Then in the third phase;
- The system extract relations between the term ("traitement") and the medical entities "*traumatisme du rachi lombaire*", "*traumatisme lombaire*", "*rachi lombaire*", "*fracture de L01*", "*fracture de L02*"; ("Lumbar spinal trauma", "lumbar trauma", "lumbar spinal", "fracture L01", "fracture L02"); this term ("traitement") will be transformed to a relation as "traite" as follow: traite (disease, X) where X will be found from the external resource.

  In our example : traite ("traumatisme du rachi lombaire", X); (traite ("lumbar spinal trauma", X)), traite ("traumatisme lombaire", X); (traite ("lumbar trauma", X)), traite ("rachi lombaire", X); (traite ("lumbar spinal", X)), traite ("fracture de L01", X); traite ("fracture L01", X), traite ("fracture de L02", X); traite ("fracture L02", X); by using the external resource ; X = "plaque vissé", "corset", "corset bivalve"; (X="screwed plaque", "corset", "bivalve corset").
- Finally, the query will be expanded to include : "plaque vissé", "corset", "corset bivalve", "traumatisme du rachi lombaire", "traumatisme lombaire", "rachi lombaire", "fracture de

L01", "fracture de L02"; ("screwed plaque", "corset", "bivalve corset", "Lumbar spinal trauma", "lumbar trauma", "lumbar spinal", "fracture L01", "fracture L02").

Semantic relation extraction with Boolean query reformulation

In this method we are interested in reformulating the user's query to Boolean expression; in the other methods, we expanded the initial query by only adding medical entities to the original query. Boolean reformulation of a user's query method suggests the use of Boolean operators. For example: if we take the previous example from the semantic relation extraction method. The expanded query will be reformulated as the Boolean expression: [("plaque vise" OR "corset" OR "corset bivalve") AND ("traumatisme du rachi lombaire" OR "traumatisme lombaire" OR "rachi lombaire" OR "fracture de L01" OR "fracture de L02")];[("screwed plaque" OR "corset" OR "bivalve corset") AND ("Lumbar spinal trauma" OR "lumbar trauma" OR "lumbar spinal" OR "fracture L01" OR "fracture L02")].

In general, the connector 'AND' is used to link medical entities that have different types. Unlike, the connector 'OR' is used to link medical entities of the same types.

EXPERIMENTS AND RESULTS

In this section, we describe the dataset and the metrics used to test our approach experimentally and discuss the obtained results. For the evaluation of our expansion methods and results comparison, we split our experimentation to different Search types; T1, T2, T3, T4 where it will be explained below. We will study them individually in order to facilitate comparing metrics to improve the retrieval performance. The search types are cited as follow:

- Test 1 (T1): simple search; searching with no expansion.
- Test 2 (T2): searching using query expansion; user's query will be expanded using synonyms and hyponyms.
- Test 3 (T3): searching by expanding query; user's query will expanded using synonyms, hyponyms and semantic relations in the context of the query.
- Test 4 (T4): searching using query expansion; user's query will expanded using synonyms, hyponyms and semantic relations with the Boolean reformulation of the query.

Test Collection

We collect over than 200 French clinical reports from general medicine at the Chlef hospital (Algeria). We choose only orthopedic clinical reports to do the test and to use a specific terminology. We also used a set of medical queries containing medical entities provided by doctors for evaluation. The external resource we used, contain medical entities and semantic relations extracted from clinical reports described in the section above. We consulted different medical websites to enriching this resource including more synonyms and hyponymy for disease[1,3,4] and even for other medical entities type[5,6,7]. And with the help of the doctors, we could validate and store more than 140 medical entities and over than 65 semantic relations.

Performance measures

Standard metrics for evaluating the effectiveness of each strategy are used; these measures are computed as follow:

- Recall is the fraction of the documents that are relevant to the query and those that are successfully retrieved.
- Precision is the fraction of retrieved documents that are relevant to the user's information need.
- F-measure (FM): the weighted harmonic mean of precision and recall, the traditional F-measure is:

$$\mathrm{FM} = \frac{2*\mathrm{precision}*\mathrm{recall}}{(\mathrm{presicion}+\mathrm{recall})} \qquad (1)$$

- Mean Average Precision: is the average precision across multiple queries/ranking.

$$\mathrm{MAP} = \frac{\sum_{q=1}^{Q} \mathrm{AveP}(q)}{Q} \qquad (2)$$

- R-precision: R-precision is the precision at R, where R is the number of relevant documents for query Q. so, if there are r relevant documents among the top-R retrieved documents;

$$\mathrm{R} - \mathrm{precision} = \frac{r}{R} \qquad (3)$$

Results

Table 4 shows the different values of the average recall, precision; F-measure, MAP and R-precision obtained by the system without and with using of the different expansion methods. We calculated for each query the recall, precision; F-measure, MAP and R-precision and then we calculated the average of the measures for all queries. We also adopted the P@10 metric which is the official measure used in search engines, P@10 denote the proportion of relevant documents in the top 10 documents in the returned list for a query request. We calculate the rate of improvement compared to the baseline (T1) and it is shown in table 4.

Metrics\methods	T1	T2	T3	T4
Recall	0.725	0.919 (+26.76%)	0.928 (+28%)	0.918 (+26.62%)
Precision	0.794	0.778 (-2.01%)	0.753 (-5.16)	**0.975 (+22.79%)**
FM	0.698	0.791 (+13.32%)	0.774 (+10.88)	0.944 (+35.24%)
MAP	0.882	0.907 (+2.83%)	0.895 (+1.47)	**0.976 (+10.65%)**
R-precision	0.668	0.833 (+24.70)	0.815 (+22%)	0.912 (+35.52%)
P@10	0.761	0.910 (+19.58%)	0.920 (+20.89%)	0.960 (+24.15%)

Table 4. The average recall/precision; F-measure / MAP, R-precision and P@10 obtained using the different methods and the rate improvement compare to T1.

Regarding the results obtained and summarized in Table 4, we can see that there is an improvement in recall compared with the baseline (T1), as well as in different expansion strategies (T2, T3, T4) there is an improvement in MAP and R-precision. The increase of recall means that there is an augmentation in the number of relevant documents retrieved by each method. In other words, we conclude that whenever the query contains more medical entities the number of relevant documents is increasing. So, the use of the external resource in query expansion improves the recall in medical information retrieval. This seems clear in T3, the rate of improvement is +28% than the baseline which combines synonyms and semantic relations.
In the other hand, the precision of the baseline looks better than T2 and T3, there is no improvement in term of precision in T2 and T3, and this means that there are a lot of irrelevant documents retrieved by the system compared to the baseline. This allowed as concluding that the query expansion techniques T2, T3 do not improve the precision in medical information retrieval, unlike T4 shows improvement in the precision rate of +22.79% compare to the baseline, it shows also an improvement in MAP, R-precision and P@10 compare to T2 and T3. So the use of the external resource with Boolean reformulation of the query improves the precision in the medical information retrieval.

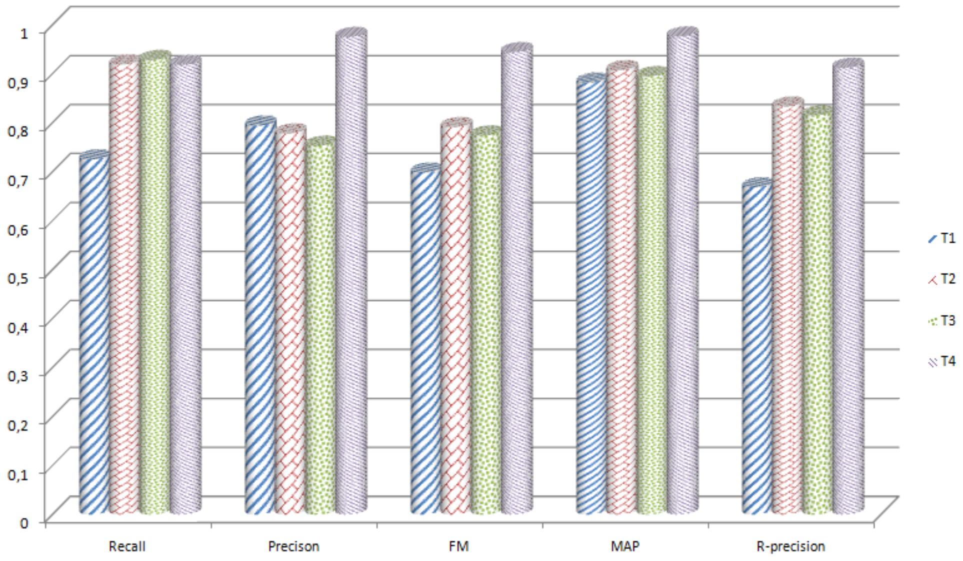


Figure 6. Performance obtained for each method using the average of recall, precision, FM, MAP, R-precision

Regarding the MAP and R-precision, we can observe from figure 6 that T4 outperforms the baseline and both techniques T2 and T3.

In other works where they used an external resource like in (khadim et al., 2014) and (yangyang et al., 2017); they evaluated the retrieval performance using metrics such as P@10. A comparison with these approaches cannot take place because the corpus of evaluation is totally different and even the language of the corpus. This is the reason that led us to create our own baseline (T1) to be able to compare it with our approach. We compare also our approach (T2) and (T3).

	T1	T2	T3	T4
0.0	0.933	1.000	1.000	1.000
0.1	0.967	1.000	1.000	1.000
0.2	0.908	0.945	1.000	0.978
0.3	0.920	0.922	0.967	0.980
0.4	0.891	0.907	0.973	0.970
0.5	0.893	0.910	0.932	0.974
0.6	0.866	0.898	0.877	0.977
0.7	0.870	0.904	0.872	0.977
0.8	0.872	0.884	0.879	0.977
0.9	0.852	0.880	0.853	0.977
1.0	0.853	0.849	0.800	0.977

Table 5. The average precision obtained in different methods.

Table 5 shows the results obtained by using the three strategies and even baseline. Precision/recall are presented showing the interpolated average precision at eleven standard recall levels.

Figure 7 shows correlation precision/recall at 11 points; we can compare the techniques and see the importance of query expansion with the Boolean reformulation of the query.

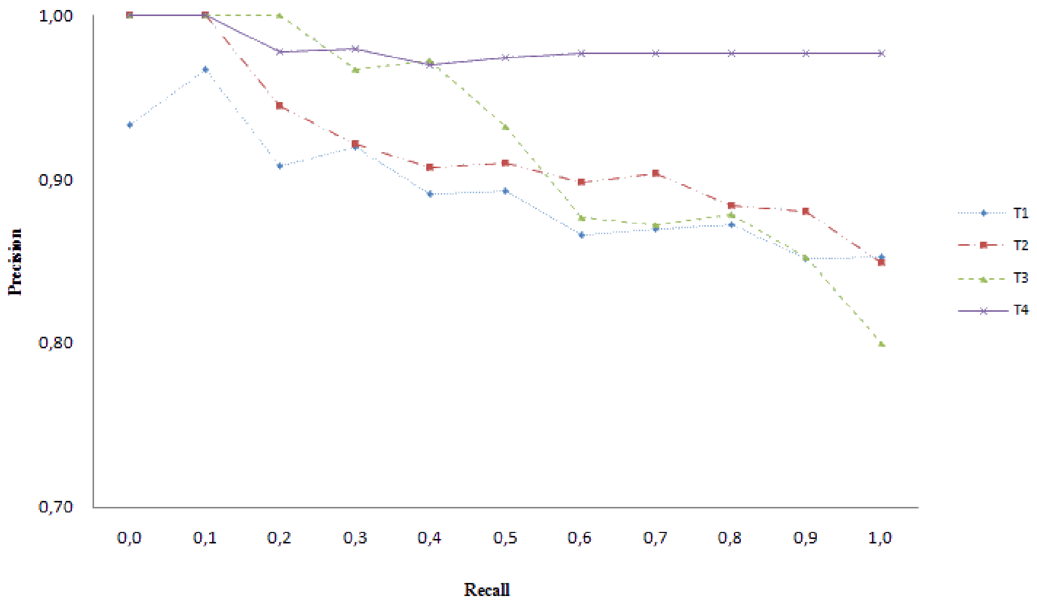


Figure 7. recall/precision curves

The experimental results in Figure 7 show that all query expansion methods achieve improvement; comparing to the baseline T1. T2 which uses expansion with synonyms and descendants improves retrieval in eight points of recall (0.0 to 0.2 and 0.4 to 0.7 and 0.9), and the performance degrades at three points (0.3, 0.8 and 1.0). Similarly, T3 query expansion using synonyms and semantic relations in the context of the query improves retrieval at seven points (0.0 to 0.6), and degrade performance at four points (0.7 to 1.0). T4 query expansion using the Boolean reformulation can be seen in recall/precision curves, it improves retrieval at all points of recall.

CONCLUSION

In this paper, we proposed a new method founded on medical entities and semantic relations extraction according to their nature and their context in the query. With the use of an external resource which is automatically constructed from extracting information from medical reports dedicated to the specific medical domain, the results of our approach are encouraging.

The proposed query expansion method with Boolean reformulation outperforms the baseline and obtained promising results, it gives a good precision compared to other approaches using external resources. However, this approach has some limits: First, We need to test this approach in a large set of French medical reports, the corpus of thousands of documents for the medical retrieval process. Second, The medical ontology construction is based on information extraction from CRs; and it does not contain synonyms for all medical entities, so it does not cover the entire medical domain; the method used for the extraction is rule-based method, this method gives very good results, however, it involves a great human effort and a considerable time for data analysis and rule writing. It is time-consuming during development. Third, we should automatically enrich our medical ontology with more synonyms, hyponyms and other semantic relations from existing medical ontologies. For these limits, we proposed as future work to include experiments in larger test collections; a corpus of thousands of French medical reports. And we will develop an approach that uses a medical existing controlled ontology such MeSH or UMLS to expand the queries and then compare it with our current approach. For more medical terms and synonyms in the ontology we may use a machine learning method that may benefit more in the extraction of information. We plan to enrich the external resource (French medical ontology) with reusing standard existing ontologies, for instance, CIM-10, CCAM, and SNOMED int.

NOTES

1. http://www.hetop.org/hetop/
2. https://lucene.apache.org
3. http://www.doctoralia.fr/maladies
4. https://fr.wikipedia.org/wiki/Colonne_vertébrale
5. http://www.e-sante.fr
6. http://www.vulgaris-medical.com
7. http://www.chirurgie-orthopedie-chanzy.com/traumatologie

REFERENCES

Abbache, A., Barigou, F., Belkredim, F., Belalem, G. (2014). The use of Arabic WordNet in Arabic Information Retrieval. International journal of information retrieval research, 4(3): 54-65.

Abbache, A., Meziane, F., Belalem, G., Belkredim, F. (2016). Arabic Query Expansion Using WordNet and Association Rules. International Journal of Intelligent Information Technologies (IJIIT) 12(3): 1-14.

Audeh. B., Baune. P., & Beigbeder. M. (2014). Exploring Query Reformulation for Named Entity. SAC'14March 24-28, Gyeongju, Korea. ACM 978-1-4503-2469-4/14/03.

Barigou, F., Beldjilali, B., & Atmani, B. (2012). Using a cellular automaton to extract medical information from clinical Reports. Journal of information processing system, 8(1): 67–84.

Ben Abacha, A., & Zweigenbaum, P. (2011). Automatic extraction of semantic relations between medical entities: a rule-based approach. Journal of Biomedical Semantics. 2(Suppl 5): S4, DOI: 10.1186/2041-1480-2-S5-S4.

Ben abacha, A., & Zweigenbaum, P. (2015). MEANS: a Medical Question-Answering System Combining NLP Techniques and Semantic Web Technologies. Information Processing & Management Journal, Elsevier, 51(5): 570-594.

Bentricia R, Zidat S, Farhi M. (2017). Extracting semantic relations from the Quranic Arabic based on Arabic conjunctive pattern. Journal of king saud university- computer and Information sciences (article in press).

Bhatnagar, P., & Pareek, N. (2014). Improving pseudo-relevance feedback based query expansion using genetic-fuzzy approach and semantic similarity notion. Journal of information science, 40(4): 523-537.

Bilel E. Ibrahim B. Oussama B. K. Fabrice E. Narjès B. B. S. (2011). Towards a Possibilistic Information Retrieval System Using Semantic Query Expansion.International Journal of Intelligent Information Technologies, 7(4), 1–25. 10.4018/jiit.2011100101

Buizza, P. (2011). Indexing concepts and or named entities. JLIS. IT. Italian Journal of Library and Information Science, 2(2) DOI 10.4403/jlis.it-4707.

Carpineto, C., & Romano. G.(2012). A survey of automatic query expansion in information retrieval. ACM Computing Surveys, 44(1) :1.

Charlet, J., Declerck, G., Dhombres, F., Gayet, P., Miroux, P., Vandenbussche, P. (2012). Construire une ontologie médicale pour la recherche d'information : problématiques terminologiques et de modélisation. 23es journées francophones d'ingénierie des connaissances, Paris, France, pp.33-48. (in French)

Chen., Y, Lu., H, Shapiro., L. Ravensara S.T, Li., L. (2016). An approach to semantic query expansion system based on Hepatitis ontology. Journal of Biological Research-Thessaloniki. 23(Suppl 1):S11 DOI 10.1186/s40709-016-0044-9.

Colace, F., Santo, M., Greco, L., &Napoletano, P. (2015). Improving relevance feedback-based query expansion by the use of a weighted word pairs approach. Journal of the Association for Information Science and Technology (JASIST); 66(11): 2223-2234. DOI: 10.1002/asi.23331.

Denis., E. C, Wasito., I. (2017). Automatic ontology construction using text corpora and ontology design patterns (ODPS) in Alzheimer's disease. Journal of a science and information. (10)2, 59-66.

Díaz-Galiano, M.C., Martín-Valdivia, M.T.,& Ureña-López, L.A. (2009): Query expansion with a medical ontology to improve a multimodal information retrieval system. Computers in Biology and Medicine; 39(4): 396-403.

Dipasree, P., Mandar M, & Samar B. (2015). Exploring Query Categorisation for Query Expansion: A Study. CoRR, (abs/1509.05567).

Embarek, M., & Ferret, O. (2012). Esculape: Un système de question-réponse dans le domaine médical fondé sur l'extraction de relations. TAL, 53(1): 69–99. (in French)

Ghoulam, A., Barigou, F., & Belalem, G. (2015a). Information Extraction in the Medical Domain. Journal of Information Technology Research, 8(2): 1-15.

Ghoulam, A., Barigou, F., Belalem, G., & Meziane, F. (2015b). Using Local Grammar for Entity Extraction from Clinical Reports. International Journal of Artificial Intelligence and Interactive Multimedia, 3(3): 16-24. DOI: 10.9781/ijimai.2015.332.

Gangemi, A., Pisanelli, D., Steve, G. (1999). An Overview of the ONIONS Project: Applying Ontologies to the Integration of Medical Terminologies. Data Knowl. Eng. 31(2): 183-220.

Hoffart, J., Yosef, M, A., Bordino, I., Furstenau, H., Pinkal, M., Spaniol, M., Teneva, B., Thater, S., Weikum, G. (2011). Robust disambiguation of named entities in text. In Proceedings of the conference on empirical methods in natural language processing (EMNLP'11). pages 782–792, Edinburgh, Scotland, UK, July 27–31.

Hersh, W., Hickam, D. (1995). Information retrieval in medicine: The SAPHIRE Experience. Journal of the American society for information science 46(10): 743-747.

Jing, Y. & Croft. W.B. (1994).An association thesaurus for information retrieval. In Proceedings of the Intelligent Multimedia Information Retrieval Systems (RIAO '94, New York, NY), 1994, pages 146–160.

Jovic, A., Marin, P., Dragan, G. (2007).ontologies in medical knowledge representation. Proceedings of the 29th Int. Conf. on Information Technology Interfaces, pages 535-540, June 25-28 cavtat Croatia.

Khadim, Dramé., Fleur, Mougin., Galo. Diallo. (2014).Query expansion using external resources for improving information retrieval in the biomedical domain. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2014). Vol-1180:189-194.

Michael., R.T.F, Priya., G, Ravichandran., K.S, Gayathri., M, Uma., R. (2016). Ontology-Based E-Healthcare Information Retrieval System: A Semantic Approach. International Journal on Recent and Innovation Trends in Computing and Communication. Volume: 4 Issue: 4. ISSN: 2321-8169, 365 - 369.

Mohameth, F., Sylvie, R., Jacky, M. (2012). OBIRS-feedback, une méthode de reformulation utilisant une ontologie de domaine. Conférence en Recherche d'Information et Applications, CORIA 2012, Mar 2012, Bordeaux, France. pp.135-150, 2012. (in French)

Nadeau, D., & Sekine, S., (2007). A survey of named entity recognition and classification. Journal of linguistic investigations, 30(1): 3-26.

Picariello A. Rinaldi A. M. (2007). User relevance feedback in semantic information retrieval.International Journal of Intelligent Information Technologies, 3(2), 36–50. 10.4018/jiit.2007040103.

Prabath., C. A, Saluka., R. K. (2012). Ontology-based information extraction for disease intelligence. International Journal of Research in Computer Science eISSN 2249-8265 Volume 2 Issue 6 (2012) pp 7-19.

Samwald., M, Feimuth., R, Luciano., JS, Lin., S, Powers., RL, Marshall., MS, Adlassing., KP, Dumontier., M, Boyce., RD. (2013). An RDF/OWL knowledge base for query answering and decision support.

Suresh., G, Zayara., G. (2014). Concept relation extraction using naive Bayes classifier for ontology-based question answering systems. Journal of king Saudi university- computer and Information sciences (2015)27, 13-24.

Yangyang., K, Jianqiang., Li, Jijiang., Y, Qing., W, Zhihua., S. (2017). Semantic Analysis for Enhanced Medical Retrieval. IEEE International Conference on Systems, Man, and Cybernetics (SMC) Banff Center, Banff, Canada.

Zulkarnain N., Meziane, F., Croft, G. (2016). A Methodology for Biomedical Ontology Reuse. 21st international conference on applications of natural language to information systems, NLDB, Salford, UK.