Je-LKS
Journal of e-Learning and Knowledge Society
The Italian e-Learning Association Journal
PEER REVIEWED PAPERS
NEW TRENDS, CHALLENGES AND PERSPECTIVES ON HEALTHCARE COGNITIVE COMPUTING: FROM INFORMATION EXTRACTION TO HEALTHCARE ANALYTICS
Vol. 14, n.1, 2018
ISSN: 1826-6223 | eISSN: 1971-8829

HANDLING NEGATION TO IMPROVE INFORMATION RETRIEVAL FROM FRENCH CLINICAL REPORTS

Baya Naouel Barigou
Fatiha Barigou
Baghdad Atmani

Computer Science Department, University Of Oran 1 Ahmed Ben Bella, Algeria
barigounaouel@gmail.com, fatbarigou@gmail.com, atmani.baghdad@gmail.com

Keywords: French clinical reports, medical information retrieval, negation identification, natural language processing.

The aim of this work is to develop a framework to cope with the negative context in French clinical reports and assess the effect of negation identification on the performance of medical information retrieval. The proposed work significantly improves the performance of information retrieval done on French clinical reports where the precision improves by 10%.

for citations:
Barigou B., Barigou F., Atmani B. (2018), *Handling Negation to Improve Information Retrieval from French Clinical Reports,* Journal of e-Learning and Knowledge Society, v.14, n.1, 11-31. ISSN: 1826-6223, e-ISSN:1971-8829
DOI: 10.20368/1971-8829/1455

1 Introduction

Electronic medical reports (EMRs) generally written in free text are considered a valuable source of clinical data; These reports are often consulted to verify a diagnosis by looking for similar cases and also see the treatment that has been prescribed. Thus, information retrieval (IR) from EMR helps clinicians identifying those reports relevant to their information need and allows them making the best clinical decision at the point of care. However, development of an IR system has always been a challenge in the medical field. In fact, processing medical documents put forth greater difficulty due to the inherent language problems like synonymy, polysemy, negation and context. In the perspective of information retrieval, this work deals with the problem of negation in French clinical reports.

Traditional keywords-based IR has many shortcomings, especially in the medical domain where the quality of the results can be decreased. Suppose a physician wants to identify a set of patients with *bronchiolitis*. The system can run a keyword search like *bronchiolitis,* which would identify documents containing the term *bronchiolitis* as given in expression 1-a. This, however, will return also documents that report a patient without *bronchiolitis*, through expressions like those in (1 b-c):

1	a	« Le début remonte à l'âge de 3 mois marqué par l'apparition d'une 1ère épisode de bronchiolite traité par Célestène-salbutamol» (The beginning dates back to the age of 3 months marked by the appearance of the first episode of bronchiolitis treated with Celestene-salbutamol)
	b	« L'enfant ne présente aucun signe de bronchiolite» ("The child has no signs of bronchiolitis")
	c	« Bronchiolite négative» ("bronchiolitis history is negative")

This inappropriate use of retrieval based on keyword search results in detriments of system precision. To identify only patients with *bronchiolitis,* the system needs to filter out expressions that do not show that the patient has a *bronchiolitis.*

Negation is a complex expressive linguistic phenomenon studied in linguistics and philosophy (Morante & Sporleder, 2012). Negation changes the meaning of sentences by denying or rejecting a statement, transforming a positive sentence into a negative sentence.

Today, negation detection is not only an emerging task in natural language processing (NLP) but also a relevant task in various applications such as Information Retrieval (IR), opinion mining and sentiment analysis.

Negation has been widely studied in medical documents. However, most of the works that have addressed this problem concern the English language. There are works that have been concerned only with identification of negation (Chapman *et al.*, 2001, Mutalik *et al.*, 2001), other works have analyzed the effect of negation during the search for information (Limsopatham *et al.*, 2012, Koopman *et al.*, 2010). In the case of the French language, less work related to medical information retrieval has focused on negated terms found in EMR. Therefore, this research paper aims to develop a complete IR system that supports negation in French EMR to reduce noise and then improve the quality of research. The developed system handles negation during reports indexing step and also during research step.

2 Background

This section summarizes the most closely developments related to our work.

2.1 Negation in the French language

Long before invention of computer science, several studies were conducted to deal with negation from a linguistic point of view: syntactic, lexical, and even semantic. However, automatic processing of negation is a newly emerging research area. In linguistics, the negation is opposed to the affirmation. And whatever the language, the negation is always marked, that is to say, that an unmarked assertion is considered, by default, as affirmative. The linguistic status of negation is therefore not equivalent to that of affirmation. The grammatical nature of the morphemes used to express negation varies according to the languages and context of enunciation. For the French language, one can mention: adverbs (e.g. ne... pas, non), adjectives (e.g. aucun, nul), verbs (e.g. ignorer: do not know), pronouns (e.g., rien (nothing), personne (nobody)), prepositions (e.g., sans (without), sauf (except)), coordinating conjunctions (e.g., ni), conjunctions of subordination (e.g. sans (without)), prefixes (eg im, mé, ana) and substantives with an intrinsic negative semantic value (e.g., carence (deficiency)); Table 1 show some forms of negation for the French language.

Table 1
DIFFERENT FORMS OF NEGATION

Type	Description	Examples
Grammatical negation (GN)	GN is expressed by grammatical structures : sans, en l'absence de, pas de, manque de, aucun...(without, in the absence of, no, lack of, none...)	"le patient n'a pas de diabète" (The patient does not have diabetes) "aucun signe de pneumonie" (No signs of pneumonia) "absence d'anésthésie" (There is no anesthesia)
Lexical negation	Several words carry the notion of negation by using prefixes Some verbs can be equivalent to a negation	"Mécontent" (unhappy), "analphabète" (illeterate) "dans ce cas, on supprime le médicament" (in this case no drugs)
Semantic negation	Interpretation of the sentence has a negative meaning	"L'examen clinique est négatif" (Clinical examination is negative) "le sirop pour la toux est contre-indiqué" (cough syrup is contraindicated)

2.2 Negation in the medical field

In the medical field, negation is a common practice when writing medical reports. Doctors use negation to express absence of signs, symptoms, or negative results of a medical test. According to (Ballesteros *et al.*, 2012), about 95% of radiology reports contain a negation of the form "*no sign of*" (in French "aucun signe de") or "*rupture of*" (In French "rupture de") or an expression in this sense. Further, doctors are not limited to the syntax of negation as defined in natural language (French for our case), they use also others expressions not known in the syntax of the French language. In Table 2 we give an example of four medical expressions that have the same semantic interpretation (i.e. negative interpretation).

Table 2
EXAMPLE OF NEGATIVE EXPRESSION USED IN FRENCH MEDICAL REPORTS

Medical expression	Interpretation
CRP[1]: (-)	Le test CRP est négatif (CRP test is negative)
CRP: négative.	
CRP négatif.	
CRP -	

Scope of negation

The question of negation scope in a sentence is critical for information

[1] C-reactive protein

retrieval. A distinction is made between the sentence negation, or phrastic, or total negation, and the constituent or partial negation. Thus, depending on the structure of the sentence, the scope would be:

Nominal phrase (NP) subject + intransitive verb→ the verb is denied. (e.g.” L’enfant ne **tousse** pas”: the *child does not cough*)

NP + ne transitive verb + pas + de + object of the verb→ the object of the verb is denied (e.g. “ l’enfant ne présente pas de **signes de pneumonie**”: *the child does not show **signs of pneumonia***)

NP + ne+ transitive verb + pas + de + object of the verb+ [, de + object of the verb]* → all the object of the verb are denied (e.g. “ l’enfant ne présente pas de **signes de pneumonie, *de diabète, de choloéstérol***”: *The child has no signs of pneumonia, diabetes, cholesterol)*

2.3 Related work

The previous work has mainly focused on detecting negated terms from medical documents to remove them before indexing documents. In addition, most of the proposed tools are designed for English-language medical texts. For example, Chapman *et al.* (2001) proposed a rule-based algorithm called NegEx for detecting negated findings and diseases in English radiology reports. Based on regular expressions, the principle of NegEx is: negation triggers (also called negation signs) are annotated in the sentence and then scope of these triggers are defined. Evaluation of NegEx uses 1000 sentences taken from discharge summaries and indexed by physicians. They obtained a specificity of 94.5% and a sensitivity of 77.8% percent. Considering the impressive results of this tool and its simplicity, it has been used in several NLP applications (Mitchell *et al.*, 2004, Meystre & Haug, 2006) and adapted to several languages like Swedish (Skeppstedt, 2011), French (Deleger & Grouin. 2012), Spanish (Stricker, Iacobacci & Cotik, 2015) and Germany (Cotik *et al.*, 2016). Similarly, Mutalik *et al.* (2001) proposed Negfinder a tool for detecting negations in discharge summaries and surgical notes. After replacement of the concepts by their corresponding UMLS terms, negation signs are identified and then a context-free grammar, known as Look-Ahead Left-to-right Rightmost-derivation associates these markers with UMLS concepts. The specificity of Negfinder was 97.7 percent, and the sensitivity was 95.3 percent.

Agrawal *et al.* (2010) present a conditional random field model, designed to detect negation cues and their respective scopes. The authors used the publicly available BioScope corpus[2] to train the proposed model. Experiments achieved an overall F-measure of 98% for detecting cues and 95% for detecting scopes.

An empirical analysis of the effect of negation on the search for information

[2] http://www.inf.u-szeged.hu/rgai/bioscope

is carried out in (Koopman *et al.*, 2010). The result of this study shows that negation has no major impact on research. But, it must be pointed out that the proposed algorithm does not consider complex forms of negation, their tool is limited to identifying simple forms such as "no", "negative" and "no". The study by Limsopatham *et al.* (2012) gets different results; unlike the results of Koopman *et al.* (2010), the effectiveness of IR has been improved. The authors propose NegFlag which is based on NegEX to detect negations. To avoid incorrectly returning a document, in which a given positive query term appears in a negative context, they propose to register a term negated by concatenating the prefix "no" to it. NegFlag is evaluated on 34 topics from the TREC 2011 Medical Record track. Compared to a baseline system, which does not consider negative context, this approach yielded a 2.9% relative improvement in P@10.

The authors in (Díaz, 2013), modeled a negation detection system by combining two classifiers in series, the first classifier determines whether a word of a sentence is a sign of negation or not and its position in the sentence. For sentences that contain a sign of negation, the second classifier determines whether the rest of the words in these sentences are in the negative part or not. Naïve Bayes (NB), C4.5, Support Vector Machine (SVM) were used in the experiment with a collection of medical documents from the corpus Bioscope. The author did not specify the values of the results but indicated that the results were good. The work of (Costumero *et al.*, 2014) is an adaptation of the NegEx algorithm (Chapman *et al.*, 2001) for the Spanish language. The results obtained for the detection of negation were weak compared to those of English texts. The authors justify this weakness by the difference between the two languages (English and Spanish) and state that a simple translation of the rules is not enough. Hence, a thorough analysis of Spanish grammar is needed to improve the performance of the Spanish version of the NegEx algorithm. Another similar work carried out by Stricker, Iacobacci and Cotik (2015) adapted NegEx for the Spanish Language. In this study, NexEg shows to perform better.

Mehrabi *et al.* (2015) propose DEEPEN to detect negation using NegEx and by taking into account the dependency relationship between negation words and concepts within a sentence. Their aim was to decrease NegEx's false positives. DEEPEN was tested using EMR data from Indiana University (IU) and it was also evaluated on Mayo Clinic dataset. Recently, Elazhary (2017) proposed a rule-based tool called NegMiner to address some of the shortcomings of the NegEx algorithm (Chapman *et al.*, 2001). Unlike NegEx, NegMiner can deal with longer sentences containing contiguous and multiple negations by exploiting some basic syntactic and semantic information. Experimental results have shown the superiority of the mining results of the NegMiner in comparison to the simulated NegEx algorithm (Chapman *et al.*, 2001).

3 Proposed work

A complete information retrieval system is proposed below. As shown in Fig.1, it consists of two components; the index-building component and the document retrieval component.

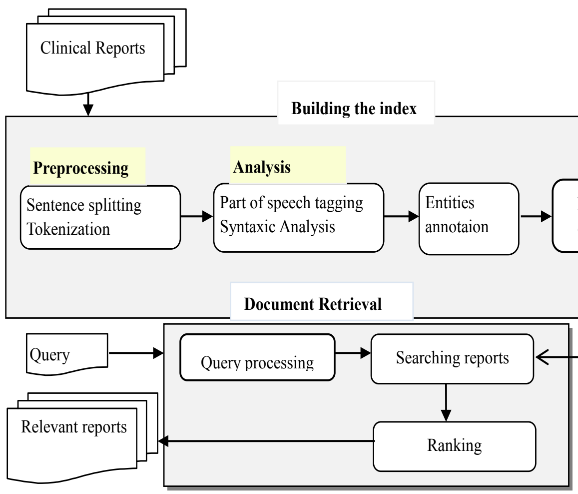

Fig.1 - Architecture of the proposed system

3.1 The index building

As indicated in Fig.1, the whole pipeline system works in five steps. We give the description of each step:

Preprocessing

The pre-processing step includes sentence boundary detection and tokenization. Decomposition of a report into a set of sentences is an important step for the detection of negation since our approach is based on the syntactic analysis of the sentences. The decomposition is mainly based on punctuation such as the dot "." the comma "," except that the majority of the sentences of the reports used in this study did not end with the point ". ", which required us to correct these reports before moving on to segmentation.

Linguistic analysis

This step includes part-of-speech (POS) tagging, lemmatization, and syntactic parsing. The Framework StanfordNLP version 3.6.0 is adopted where the French tagger library is used for POS tagging. Next, the results from POS tagging are fed to the Stanford parser to generate the syntactic tree.

Entities annotation

After linguistic analysis, terms annotation is carried out to distinguish between medical and non-medical terms. We need this annotation to weight the query terms during the research phase. The medical terms are more informative; as a result, they must have a higher weight than ordinary terms.

Medical entity identification is done using dictionary lookup. We compiled a list of commonly used entities (names of diseases, drugs, and symptom) from web sources like http://dictionnaire.doctissimo.fr and https://www.vidal.fr/ using the xidel[3] tool.

This resource is therefore used to annotate terms present in EMRs as medical terms if they are found in this resource, ordinary terms otherwise. Consider the following sentence as an example:

2A *Pas de notion de tuberculose pulmonaire mais présence de douleur thoracique (“No pulmonary tuberculosis but presence of chest pain “)*

pas :negation cue (“no” in english) considered as a stop word, it is not stored in the index
de : (“of”) stop word not stored in the index
notion: ordinary term
tuberculose: medical term
pulmonaire: medical term

In the above sentence of example 2-A, “*tuberculose*” (“*tuberculosis”*) is annotated as a medical term.

Negation detection and scope finding

Detection of negation and scope finding is typically divided into two subtasks. That is, first the trigger words of negation (negation cue) are detected and, thereafter, the task of determining the scope is carried out. Thus, a pattern-searching algorithm aimed to detect every word related to a negation cue analyzes the resulting syntactic tree. This step is performed based on a supplied list of models representing a negation format. A model of negation includes the syntactic expression of a sentence containing a sign of negation (e.g., “ne”, “ni”,

[3] https://www.phpclasses.org/package/9136-PHP-Extract-information-from-HTML-using-the-Xidel-tool.html

"pas", "rien", "aucun", etc) and indicates the scope of this negation (the terms denied) in this sentence.

Detection of negation involves two steps. Candidate models are first identified; a vocabulary of trigger words of negation is compiled and words in the sentence are compared to this vocabulary in order to determine whether they are negation cues. If so, only models containing sign of negation similar to that of the sentence being analyzed are selected. In the second step, a match between the syntactic tree of the sentence and each candidate model is realized. This step makes it possible to identify the appropriate model that goes with the syntactic structure of the sentence.

Once the model corresponding to the sentence is identified, comes after the identification of negated terms (scope of negation sign). Once again, the two trees are compared in order to extract from the graph of the sentence the negated terms.

According to the French grammar, the scope of a negation trigger starts from the first word following the negation trigger and extends to the end of the sentence. Nevertheless, there are exceptions where this rule is not respected; for example, the use of words (such as "mais" (but) in example 2) causes a premature termination of the negation scope. In addition, the scope of some negation triggers is to the left of the trigger instead of the right Consider the following models as an example: [456]

3 A. (SENT (NP (ADV (**pas**[4])) (P (*)) (NC (*negation[5]*)) (AP (ADJ (*négation*)))))
B. (SENT (NP (ADV (**pas**))(P (*)) (NC (*negation*)) (PP (P (*)) (NP (*negation*))))mais NP)
C. (SENT (NP (*negation*)) (PUNC (*)) (NP (ADJ (**négatif**[6]))))

The regular expressions of example 3 are used to identify the negative diagnosis. The two first examples depict the syntactic patterns of sentences containing the negation sign "*pas*" ("not" in English). The last one contains the syntactic structure of a sentence containing the negation sign "*négatif*" ("negative" in English).

Using the model 3-B for the sentence of example 2-A, we can easily identify that the scope of "*pas*" which is a trigger word for negation, is "*notion de tuberculose pulmonaire*". So, "tuberculose" and "pulmonaire" are both denied. Note that the nominal phrase "*présence de douleur thoracique*" is not affected by negation because of the presence of the word "*mais*" ("*but*").

[4] trigger word for negation
[5] scope of negation
[6] trigger word for negation

Stop Words elimination

Unlike the usual indexing process, the step of eliminating empty words occurs after the negation detection step and not at the beginning of the process. Indeed, some signs of negation such as "ne (no), pas (not), sans (without), aucun (no) " are considered empty words, but these words play an important role in the negation detection phase. For this purpose, this filtering step is postponed after negation detection step.

Indexer

When building the index, we consider the following constraints:

1. Retrieve information from clinical reports that are written in the French language
2. Distinguish between medical terms and normal terms when ranking retrieved documents
3. Consider the terms that are denied in the weighting formulae of ranking

To deal with negation when retrieving documents as a response to a user query, the proposed index must satisfy the following conditions:

1. The index must record whether and at which position a term occurs in a document,
2. The index must record whether and at which position a negated term occurs in a document,
3. The index must record all information about a term "ti" necessary for its weighting:
    - Term Frequency (TF(ti, dj)): The total number of occurrences of the term "ti" in a document dj without considering whether it is denied or not.
    - Document Frequency (DF(ti)): The total number of documents containing term "ti".
    - Term Denied Frequency (TDF(ti, dj)): the number of times the term "ti" is negated in a document dj.
    - Document Denied Frequency (DDF(ti)): the number of documents containing the denied form of ti.

To deal with medical information retrieval, we have decided to distinguish between ordinary terms and medical terms during the retrieval. For that, in the case of noun terms, the index must record their type (medical term or ordinary term).

Thus, the core of the index consists of:

1. The dictionary represented by a hash table containing representative terms in documents. The dictionary records the root of the term, its DF, DDF, type ("M": medical or "O": Ordinary) and a pointer to its postings list.
2. The postings list: a linked list is associated with each term in the dictionary. Each posting is a docID with a list of positions, TF and TDF. When an occurrence of the term is negated, the position of this occurrence is also negated.

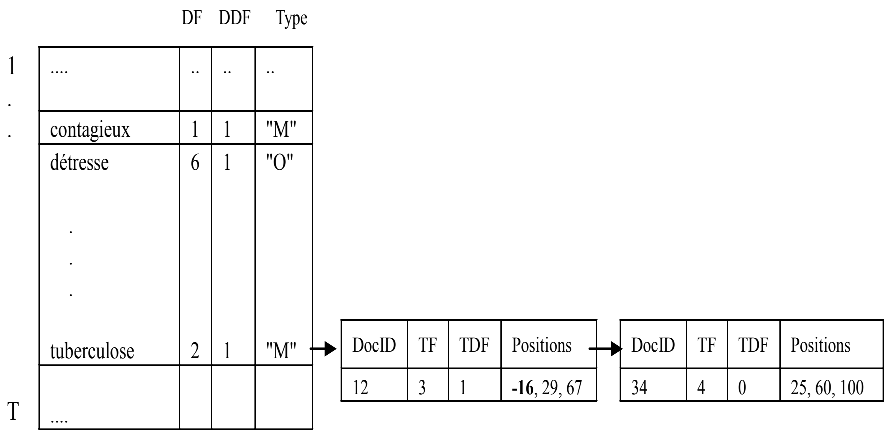

Fig.2 - Dictionary and Posting Lists structure

As an example, we have in Fig.2, the term "tuberculose" ("tuberculosis") labeled as a medical term (Type = "M"), it is present in 2 reports (DF= 2) and negated in one report (DDF=1). In the report of ID 12, this term occurs three times (TF =3) where one occurrence is negated (TDF =1). The negated occurrence is at the position 16 (the position is prefixed with minus when the occurrence is negated).

In general, the structure used for the index is crucial for the efficiency of Information Retrieval (IR) systems. To speed up the retrieval of documents, we used Bitset vectors for representing Document-Term structure (DT), Document-Term Denied structure (DTD) and Term-Document structure (TD). The details about these three vectors are given in the following section. We consider "T" as the size of the dictionary (number of representative terms) and "N" as the number of documents in a collection.

Document-Term structure

The document-term structure is a list of Bit set vectors; DT(j); j=1 to N, where the j-th Bitset vector (noted DT(j)) corresponds to the jth document in the collection (its DocId from 1 to N) and the i-th Bit of each DT(j) represents the presence/absence of term "ti" in the j-th document. It is set to 1 if "ti" is present in this document, 0 otherwise.

Document-Term Denied structure

The document-term denied structure is like the Document-Term structure but here we focus on a document containing terms that are denied. The j-th Bitset vector (noted DTD(j)) corresponds to the jth document in the collection (its DocId from 1 to N) and the i-th Bit of each DTD(j) is set to 1 if all the occurrences of the i-th term are negated in this j-th document, otherwise 0.

Term-Document structure

The term-document structure is a list of Bit set vectors; TD(i); i=1 to M, where the i-th Bitset vector (noted TD(i)) corresponds to the i-th term of the dictionary (its row number in the array from 1 to M) and The j-th Bit of each TD(i) represents the presence/absence of the term "ti" in the document "dj" (j= 1 to N). It is set to 1 if the term is present in this document, 0 otherwise.

3.2 Retrieval step

Candidate documents searching

To speed up the process of retrieving documents and consider the presence of negated terms in these documents, we used both Bit set vectors' DT and DNT to get only documents that match the query.

Also, the query "Q" is represented with a Bit set vector Q_v of size M where the i-th bit is set to 1 if the corresponding i-th index term is present in the query. It is set to 0 otherwise.

As shown in equation 1, identification of candidate documents consists in calculating for each document d_i(i = 1 to N) the Boolean vector *R(i)* using the Boolean vectors DT(i), DNT(i) and Q and the Boolean operations XOR and AND.

$$\mathrm{R(i)} = (\mathrm{DT(i)} \otimes \mathrm{DTN}) \wedge Q_v \qquad \text{eq.1}$$

According to equation 1, the document d(i) is a candidate document if $\exists\, \mathrm{j} \in [1, \mathrm{M}]/\mathrm{R(i, j)} = 1$, so documents containing the terms of the query

"Q" but which are denied are rejected by the system.

Scoring and ranking model

We used the state of the art term weighting models namely BM25 (Robertson *et al.*, 1996) to score and rank medical reports. For a given query q, the relevance score of document d, based on the BM25 term weighting model is expressed as

$$\text{Score}_{BM25}(d, q) = \sum_{t \in q} IDF(t) * \frac{TF(t,d).(k+1)}{TF(t,d)+k.\left(1-b+b.\frac{L_d}{L_{ave}}\right)} \quad \text{eq.2}$$

where

- $IDF(t) = log\left[\frac{N}{DF(t)}\right]$: $DF(t)$ is the document frequency of term *t* and *N* is the number of documents in the collection;
- $TF(T_i,d)$ is the frequency of term T_i in document *d*
- k is the term frequency influence parameter which is set to 1.2 by default;
- *b* is the document normalization influence tuning parameter which has the default value of 0.75 (Robertson *et al.*, 1996);
- L_d and L_{ave} are respectively, the length of document *d* and the average document length for the whole collection.

In this work, we review the formula and added new information to reflect the presence of the negated terms in the documents and to associate weights to query terms. The relevance of a document for a given query is usually evaluated based on the weight of the query terms. This weight depends on the frequency of occurrence of the term in the document and, possibly, its frequency of occurrence in the collection. Compared to this classical model, taking account of the negation of terms in the weighting scheme consists in attributing weight to terms by considering their presence in a positive context or a negative context. Thus, the relevance of a document no longer depends only on the frequency with which the terms of the query appear, but also on the context (negative or positive) in which these terms appear in the document.

Thus, we propose a new BM25 weighting version for classifying relevant documents. This new version supports two important points:

- It considers that a medical term is more relevant than an ordinary term
- It considers the terms denied in the documents found by the system

The revised formula is given by the equation 3:

$$Score_{Mod\_BM25}(d, q) = \sum_{t \in q} w_t * IDF(t) * \frac{(TF(t,d) - NTF(t,d)).(k+1)}{TF(t,d) + k.\left(1 - b + b.\frac{L_d}{L_{ave}}\right)} \quad \text{eq.3}$$

- *NFT(t,d)* is the number of times the query term *t* is negated in document *d*
- w_t is the weight of term *t*
- $IDF(t) = \log \frac{N - (DF(t) - NDF(t)) + 0.5}{DF(t) + 0.5}$
- *NDF(t)* is the document frequency for which the term *t* is negated

We propose an extension of the probabilistic model (Maron *et al.*, 1960; Robertson *et al.*, 1976) that evaluates the relevance of a document for a given query through two probabilities: the probability of finding relevant information and finding irrelevant information. This extension exploits the fact that query terms can be negated in a medical report and in this case, this document must not be relevant to this query. Thus, a term will not have the same importance if it appears in a negative context or positive context. Document *d* is retrieved only if at least one appearance of the query term *t* in that document is not negated.

We consider that medical terms are more informative than the ordinary terms. For example, in "*détresse respiratoire*" ("respiratory distress") the medical term "respiratoire" is the most informative term to understand the noun phrase semantics while "détresse" ("*distress*") is an ordinary term that can be used in other noun phrases like "*détresse pulmonaire*" ("*pulmonary distress*"). For that, the weight w_t in the equation 3 takes into account the type of the query term. A medical term will have a greater weight than a non-medical term..

Let Q be a query and w_i be the weight of the ith term of the query Q.

We have then:

$$\sum_1^{|q|} w_i = |NMed| * w_{NMed} + |Med| * w_{Med} = Q_{NMed} + Q_{Med} = 1 \quad \text{eq4}$$

where

- |NMed| and |Med| are respectively the number of non-medical terms and the number of medical terms in the query Q (|NMed| + |Med| =|Q|)
- w_{NMed} and w_{Med} are respectively the weights of a non-medical term and a medical term.
- Q_{NMed} and W_{Med} are respectively the quantities of information provided by non-medical terms and medical terms in the query Q.

We can have three situations according to the values of |NMed| and |Med|:

1. All the terms of the query Q are non-medical terms, i.e. |Med|= 0 and |NMed| ≥1. In this situation we have :

$$Q_{NMed} = 1 \rightarrow \; w_{NMed} = \frac{1}{|NMed|} \qquad \text{eq.5}$$

2. All the terms of the query Q are medical terms, i.e. |NMed|= 0 and |Med| ≥1. In this situation, we have :

$$Q_{Med} = 1 \rightarrow \; w_{Med} = \frac{1}{|Med|} \qquad \text{eq.6}$$

3. The query **Q** contains **|Med|** medical terms and |**NMed|** non-medical terms**, i.e.** |NMed|>0 and |Med|> 0

We calculate the weight of nom-medical terms following the equation 7

$$w_{NMed} = \frac{1}{|NMed|(1+\beta)} \qquad \text{eq.7}$$

We express the importance of the quantity of information Q_{Med} versus the quantity of information Q_{NMed} with the coefficient of importance β. To avoid assigning too much weight to non-medical terms we impose a constraint to ensure that a medical term will always have a higher weight than a non-medical term. We then have the following system of equations:

$$\begin{cases} \beta = \frac{Q_{Med}}{Q_{NMed}} = \frac{|Med| * w_{Med}}{|NMed| * w_{NMed}} \\ w_{NMed} < w_{Med} \\ \beta > 1 \end{cases} \qquad \text{eq.8}$$

The system admits a solution if $\mathbf{1} < \frac{|Med|}{|NMed|} < \beta$ so it suffices to take a value $\boldsymbol{\beta} = \frac{|Med|}{|NMed|} + \mathbf{1}$

4 Experimental study and results

We present experiments to evaluate the performance of our proposed method. Our evaluation baseline is the BM25 model. Moreover, we compare our approach with the traditional keywords-based IR.

4.1 Document collection and Queries

Using a French corpus presents some challenges. To date, the pediatric hospital of Canastel[7] uses patients' records in the paper form and there is no real electronic benchmark. In this study, the pulmonary unit gives us access to 100 anonymous French reports for the period 2010 and 2011. We correct these reports (sentences which are not syntactically well-formed, and when there is lack of punctuation), enter, and save them in a text file format (txt). We also prepare a set of 21 queries that doctors may pose when faced with signs and symptoms of a medical condition of a new case. These queries are formulated with the help of a doctor of the same service. Priority is given to queries containing terms that are found in negated context.

4.2 Evaluation measures

We measure the retrieval performance of the proposed system in the term of Precision (P), Recall (R), and Mean Average Precision (MAP). The formulas of these measures are well defined in the book of (Manning *et al.*, 2008).

Precision (P) for a query Q is defined as the fraction of retrieved documents that are relevant (RDR) within the set of documents returned for Q (*RetD)*.

$$P = \frac{RDR}{RetD} \quad \text{eq.9}$$

Recall (R) is the fraction of relevant documents that are retrieved by the system (RDR) over all relevant documents *RelD*.

$$R = \frac{RDR}{RelD} \quad \text{eq.10}$$

Mean Average Precision (MAP) provides a single-figure measure of quality across recall levels. It is given by

$$MAP = \frac{1}{|NQ|}\sum_{j=1}^{|NQ|} AP_j \quad \text{eq.11}$$

Where $|NQ|$ is the number of queries and AP_j is the average precision for a given query Q_j

$$AP_j = \frac{1}{nj}\sum_{i=1}^{nj} P@k \quad \text{eq.12}$$

nj is the number of relevant documents for Q_j and *P@k* is the precision at rank position *k of the relevant document.* We used k=5.

[7] Located at west of Algeria

To evaluate the impact of the proposed index representation we have launched several experiments considering the following objectives:

- Impact of negation detection on Boolean retrieval
- Performance of the revised weighting measure BM25

4.3 Experiments results

Impact of negation detection on Boolean retrieval

For each query Qi (i= 1, 21), we calculate precision and recall. Then an average for each measure is calculated. The results obtained are shown in Table 3. As a first result, the Boolean search with the AND operator gives the best results in precision. It is clear that when the query is expressed in several terms, the precision of the system degrades with the OR operator. In addition, the most interesting of these results is that when the negation is active, the performance of the system still improves for queries that contain terms that are negated in some reports. These are, for example, requests Q1, Q4, Q7, Q9, Q15, and Q17. In these queries, considering the AND operator, the precision is much better when the negation is taken into account. On average, when the negation is not active, precision reaches the value of 67%. Moreover, when, the negation is active, precision increases with a rate of 8.58% to reach 71.77%. Fig.3 illustrates the performance of the system in the term of precision when using negation detection. As queries, we used only those that contain at least one term that is negated in clinical reports. We can easily see that the detection of negation greatly improves the results of research.

Table 3
IMPACT OF NEGATION DETECTION ON BOOLEAN RETRIEVAL; |Q| REPRESENTS THE NUMBER OF TERMS IN QUERY Q

Negation		non-active		non-active		active		active	
Operator		OR		AND		OR		AND	
	|Q|	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
Q1	3	18,75	100,00	65,00	100,00	19,05	100,00	80,00	100,00
Q3	2	3,45	100,00	18,18	100,00	3,45	100,00	18,18	100,00
Q4	1	66,67	100,00	66,67	100,00	100,00	100,00	100,00	100,00
Q5	1	77,78	100,00	77,78	100,00	77,78	100,00	77,78	100,00
Q6	2	45,45	100,00	45,45	100,00	45,45	100,00	45,45	100,00
Q7	4	37,50	100,00	62,16	100,00	44,44	100,00	98,46	100,00
Q8	3	38,89	100,00	66,67	100,00	38,89	100,00	66,67	100,00
Q9	2	6,41	100,00	71,43	100,00	6,94	100,00	83,33	100,00

Negation		non-active		non-active		active		active	
Operator		OR		AND		OR		AND	
	|Q|	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
Q10	3	15,00	100,00	37,50	100,00	16,67	100,00	37,50	100,00
Q11	1	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00
Q12	2	11,29	100,00	53,85	100,00	11,29	100,00	53,85	100,00
Q13	2	34,78	100,00	68,57	100,00	34,78	100,00	68,57	100,00
Q14	2	16,13	96,15	66,67	96,15	16,13	96,15	66,67	96,15
Q15	2	6,67	100,00	66,67	100,00	45,00	100,00	100,00	100,00
Q16	2	12,50	100,00	12,50	100,00	25,00	100,00	25,00	100,00
Q17	2	6,15	100,00	80,00	100,00	7,41	100,00	100,00	100,00
Q18	2	5,26	100,00	33,33	100,00	5,26	100,00	33,33	100,00
Q19	2	5,88	100,00	100,00	95,83	5,88	100,00	100,00	95,83
Q20	5	3,57	100,00	100,00	100,00	3,57	100,00	100,00	100,00
Q21	1	71,43	100,00	71,43	100,00	80,65	100,00	80,65	100,00
Average		29,18	99,81	**63,19**	99,60	34,38	99,81	**71,77**	99,60

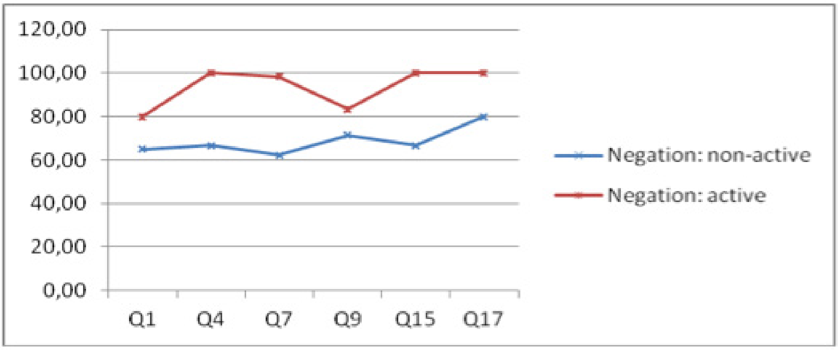

Fig. 3 - Performance of the system in term of precision when considering negation (here the AND operator is used for searching)

Measuring performance of the modified BM25

In the following experiments, we compute the average precision (AP) of each query then we compute the mean AP over the set of queries (MAP) by considering a search with the AND operator and taking into account the negation detection. The objective of these experiments is to analyze the performance of the modified BM25 measurement, which considers negated terms in reports. Results are presented in Fig.4.

Compared to BM25, it is clear that the new measure, which takes into account the terms negated in the reports, improves the quality of research. In addition, when the query includes medical terms the weighted Mod_BM25 gives the best results. As shown in Fig.4, the MAP of the modified BM25 (with/without term weighting) is the best.

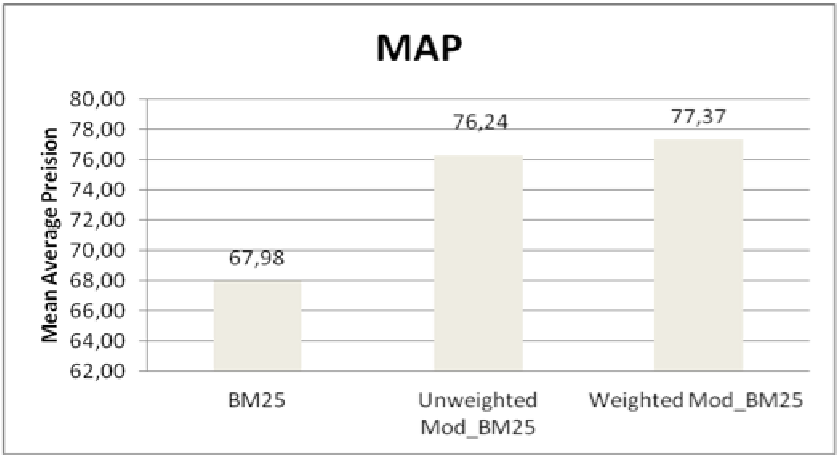

Fig.4 - The mean average precision with different term scoring

Conclusion

To improve the search quality of medical information retrieval, our solution is to introduce a negation detection mechanism to take into account this linguistic phenomenon not only during the indexing process but also during the search.

For this, we have designed and implemented an IR system based on a hybrid model (Boolean /probabilistic). The Boolean representation of the index allows to filter irrelevant documents i.e. documents containing no query terms or containing negated terms. The proposed adaptation of the BM25 formula takes into account the query terms that are negated in the reports and, finally the proposed term weighting considers the type of the query terms (medical or ordinary).

The results of our experiments, however modest they are, are still very encouraging. We achieved a significant improvement in precision compared to a system without detection of negation. These results confirm the importance of considering negation during the information retrieval process.

As prospects, there are still ways to explore and improvements to make. Therefore, several points can be considered:

- Enrich the base of negation models: A system that allows automatic generation of models from negative sentences annotated manually will ensure consistency between models and the detection of conflict situations or model redundancies because, with the evolution of the base models, the manual management of models will not be an easy task.

- Consider the structure of medical reports: medical reports have a well-defined structure that is characterized by a number of sections such as history section, clinical examination, and so on. A model that supports the search for answers at the different sections of the medical report will surely improve the accuracy of the results. This will allow the physician to search the information in well-defined sections of the document and thereby improve the accuracy of the results.

REFERENCES

Afzal, Z., Pons, E., Kang, N., Sturkenboom, M., Schuemie, M.J., Kors, J.A. (2014), *ContextD: an algorithm to identify contextual properties of medical terms in a Dutch clinical corpus BMC Bioinformatics*, 15, p. 373

Agarwal, S., Yu, H. (2010), *Biomedical negation scope detection with conditional random fields*. Journal of the American medical informatics association, 17(6):696–701.

Baldonado, M., Chang, C.K., Gravano, L., Paepcke, A. (1997), *The Stanford Digital Library Metadata Architecture*. *Int. J. Digit. Libr.* 1 108–121

Ballesteros, M., Francisco, V., Díaz, A., Herrera, J., & & Gervás. (2012), *Inferring the scope of negation in biomedical documents*. Computational Linguistics and Intelligent Text Processing, 363-375.

Bruce, K.B., Cardelli, L., Pierce, B. C. (1997), *Comparing Object Encodings*. Abadi, M., Ito, T. (eds.): Theoretical Aspects of Computer Software. Lecture Notes in Computer Science, Vol. 1281. Springer-Verlag, Berlin Heidelberg New York 415-438

Chapman, W., Bridewell, P., Hanbury, Cooper, G. F., Buchanan, B. G. (2001), *A simple algorithm for identifying negated findings and diseases in discharge summaries*. Journal of Biomedical Informatics, 34(5):301- 310.

Costumero, R., Lopez, F., Gonzalo-Martiın, C., Millan, M., & Menasalvas, E. (2014), *An Approach to Detect Negation on Medical Documents in Spanish*. Brain Informatics and Health, 366-375.

Díaz, N. P. (2013), *Detecting Negated and Uncertain Information in Biomedical and Review Texts*. The 9th International Conference on Recent Advances in Natural Language Processing, 45-50.

Elazhary, H. (2017), *NegMiner: An Automated Tool for Mining Negations from Electronic Narrative Medical Documents*. International Journal of Intelligent Systems and Applications, 9(4), pp.14-22.

Huang, Y., Lowe, H.J. (2007), *A novel hybrid approach to automated negation detection* *in clinical radiology reports*. Journal Am Med Inform Assoc, 14, pp. 304–311

Koopman B, Bruza P, Sitbon L, Lawley M (2010), *Analysis of the effect of negation on information retrieval of medical data*. Proceedings of the Fifteenth Australasian Document Computing Symposium (ADCS), Melbourne, Australia, pp 89–92

Limsopatham, N., Macdonald, C., McCreadie, R., Ounis, I. (2012), *Exploiting term dependence while handling negation in medical search*. Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 1065-1066. ACM.

Manning, D. C., Raghavan, P., and Schutze, H.(2008), *Introduction to Information Retrieval*. Cambridge Univ. Press, Cambridge

Maron, M. E, Kuhns, J. L. (1960), *On relevance, probabilistic indexing and information retrieval*. Journal of the ACM.

Mehrabi, S., Krishnan, A., Sohn, S., Roch, A.M., Schmidt, H., Kesterson, J., Beesley, C., Dexter, P., Schmidt, M., Liu, H., Palakal, M. (2015), *DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx*. Journal of Biomedical Informatics, 54:213–219.

Michalewicz, Z. (1996), *Genetic Algorithms + Data Structures = Evolution Programs*. 3rd ed. Springer-Verlag, Berlin Heidelberg New York.

Mitchell, K. J., Becicha, M. J., Bermanc, J. J., Chapman, W., Gilbertsona, J., Guptaa, D., *et al.* (2004), *Implementation and Evaluation of a Negation Tagger in a Pipeline-based System for Information Extraction from Pathology Reports*. Medinfo, 663-7.

Meystre, S., Haug, P.J. (2006), *Natural language processing to extract medical problems from electronic clinical documents: performance evaluation*. J. Biomed. Inform. 39(6), 589-599.

Morante, R., Sporleder, C. (2012), *Modality and Negation: An Introduction to* the Special Issue. Computational Linguistics 38(2), 223-260.

Mutalik P., Deshpande A., Nadkarni P. (2001), *Use of general-purpose negation detection to augment concept indexing of medical documents*. Journal of the American Medical Informatics Association, 8(6), 598.

Robertson, S, Sparck Jones, K. (1976), *Relevance weighting of search terms*. Journal of the American Society for Information Science.

Roch, A.M., Mehrabi, S., Krishnan, A., Schmidt, H.E., Kesterson, J., Beesley, C., *et al.* (2014), *Automated pancreatic cyst screening using natural language processing: a new tool in the early detection of pancreatic cancer HPB*

van Leeuwen, J. (ed.) (1995), *Computer Science Today*. Recent Trends and Developments. Lecture Notes in Computer Science, Vol. 1000. Springer-Verlag, Berlin Heidelberg New York.