

Chapter 2

Automatic Medical Image Multilingual Indexation Through a Medical Social Network

Mouhamed Gaith Ayadi, Riadh Bouslimi, Jalel Akaichi, and Hana Hedhli

1 Introduction

Online social networking is attracting more and more people in today's Internet, where users can share and consume all kinds of multimedia contents [1]. Like most people, healthcare professionals use mainstream social media networks to connect with friends and family. But almost one-third of them also join social networks focused exclusively on healthcare, where healthcare professionals can collaborate and share resources online, and patients can access more than information [2]. According to Doganay [2], patient-focused networks, often built around a particular condition or disease, give individuals and their families' supportive communities where they receive comfort, insights, and potential leads on new treatments. The data mining practices of sites like Facebook and Twitter make some patients and providers leery of posting questions or comments; and while many healthcare organizations use Facebook, Twitter, LinkedIn, Instagram, and other social tools to communicate with constituents, individuals often worry about posting information in the wrong place. By sharing data on specialized sites healthcare professionals and other users can feel safer about expressing their thoughts [2]. Franklin and Greene [3] consider that participation in the health care management can render patients longer health conscious. According to Grenier [4], the main objective behind medical networks is to foster collaboration between medical actors and to place the patient at the heart of the health system. In reality, the fact of making important decisions, related to medical images, individually, can lead the

M.G. Ayadi (✉) • R. Bouslimi • J. Akaichi

Department of Computer Sciences, ISG, BESTMOD, Tunis, Tunisia

e-mail: mouhamed.gaith.ayadi@gmail.com; bouslimi.riadh@gmail.com; jalel.akeichi@isg.rnu.tn

H. Hedhli

Emergency Department, Charles Nicolle Hospital, Tunis El Manar University, Tunis, Tunisia

e-mail: hedhli_hana@yahoo.fr

physician to make errors leading to malpractices and consequently to unexpected damages. This fact is justified by a study done by The Institute of Medicine of the National Sciences Academy¹ (IMNAS) in the USA. This institute published a study estimating that up to 98,000 hospital deaths each year can be attributed to medical malpractice [5]. In order to minimize medical errors, a medical social network, as a first contribution, destined to present patients' medical images and physicians' interpretations expressing their medical reviews and advices present the solution to support collaboration between physicians and patients. This will, obviously, help to save time and better serve the patients about their situations.

Medical images and comments present the major fields among others, meant for interaction between patients and physicians. So, an attempt to index medical images content shared through the social network site becomes an important task due to the huge number of comments expressing specialist's analysis and reviews. For this reason, an analysis approach which extracts keywords and terms from comments must be used to give an overview and a summary of what exist on comments. Furthermore, extracted terms will be used to annotate and to index images in order to facilitate the search task later, through the social network site. But, we need to take into account that existing comments can be expressed in different languages because we cannot force the social network users to use one language to present their problems (for patients) or to present analysis and recommendations (for physicians). The mechanisms, used in the indexing phase, need to be adapted to the characteristics of different languages. To overcome this problem, we will present a medical images multilingual indexation based on statistical methods and on external semantic resources. This multilingual indexation's challenge has widely taken into consideration to improve the search of images later. To improve our indexing process, we added a correction of spelling mistakes using a medical vocabulary.

The remainder of this chapter is structured as follows. Section 2 presents the background needed to follow our methodology. Section 3 describes the design of our social network. Section 4 details the proposed methodology. Section 5 shows experimental measurements, discusses the experiments, and analyzes the achieved results. Finally, Sect. 6 draws some conclusions and highlights future research directions.

2 Related Work

This section on related work begins with different medical social networks main requirements and data structure, presenting the collaboration between physicians and patients. We then study work related to the main approaches dealing with an overview of images multilingual indexation firstly and some works dealing with images indexation through social networks.

¹<http://iom.nationalacademies.org/>

2.1 Medical Social Networks

Today, social networks have the ability to connect people with just about everything. The influence of social network and those using social networks grows and changes daily, generating a profound impact on society. Furthermore, a growing majority of modern patients, particularly those with chronic conditions are seeking out social network and other online sources to acquire health information, connect with others affected by similar conditions, and play a more active role in their healthcare decisions [6]. In 2011, more than 80% of adults reported using the Internet as a resource for health care quality information and more than half of patients (57%) said they were more likely to select hospitals based on their social media presence [7]. Indeed, research shows that 81% of consumers believe that if hospitals have a strong media presence, they are likely to be more innovative than other hospitals. According to the Centers for Disease Control and Prevention (CDC), "Using social network tools has become an effective way to expand reach, foster engagement and increase access to credible, science-based health messages." [7] Medical networks have several forms which can be networks of hospital management (internal coordination and fragmentation within the hospital by specialty), resource networks (shared resources such as scanners), information networks (data collection to adjust policy information), and among many other care networks [4], which offers the members suffering from diseases an opportunity to change their lives, connect with others, and share problems. Indeed, research shows that 81% of patients believe that if hospitals have a strong media presence, they are likely to be more innovative than other hospitals.²

According to Feldman [7], hospitals are increasingly adopting the use of social network for a variety of key tasks, including: education and wellness programs, crisis communication, staff and volunteer training, employee and volunteer recruitment, information sharing, clinical trial recruitment and other research, public relations and marketing, etc. Since the beginning of 2011 alone, the growth in social network use for hospitals has been staggering. Ed Bennett, the manager of web operations at the University Medical Center in Baltimore, has been tracking hospital social media on his private site since 2008. He reports that as of October 2011, nearly 4000 social media sites were owned by US hospitals [7]. Many examples of medical social networks were presented in relation with different medical activities: SoberCircle³ is intended for alcoholics and drug addicts in need for support and encouragement by others. SparkPeople,⁴ Fitocracy,⁵ and Dacadoo⁶ do share

²<http://corp.yougov.com/healthcare/>

³<http://www.sobercircle.com>

⁴<http://www.sparkpeople.com/>

⁵<https://www.fitocracy.com/>

⁶<https://www.dacadoo.com/>

workouts and exercises in order to sustain them during weight loss. Asklepios,⁷ exclusively for Canadian doctors, is meant to exchange the best practices ever and to learn from each other. CardioSource,⁸ Cardiothoracic Surgery Network, concerns cardiothoracic surgery. Diabspac⁹ is intended for diabetics. “Parlons Cancer”¹⁰ is dedicated to cancer patients and their families. Renaloo¹¹ deals with kidney disease, dialysis, and transplantation. RxSpace¹² is dedicated to pharmacy students, pharmacists, pharmacy owners, and academia to interact with each other.

Besides, we can summarize that the top 25 health and medical social network sites are organized in three: Social networks for Doctors which offer great opportunities to confer find support and provide their own expertise, such as Sermo (<http://www.sermo.com/>), Ozmosis (<https://ozmosis.org/>), Doctor Network (<http://doctor-network.com/>), etc. Social networks for nurses which can help them to connect with others who understand what is happening in the field ask and answer questions and learn more about their profession, such as Nursing Link (<http://nursinglink.monster.com/>), Ultimate Nurse (<http://www.ultimatenurse.com/>), and Nurse Zone (<http://nursezone.com/>). And finally, social networks for all health and medical careers, such as MedicalMingle (<http://www.medicalmingle.com/>), Radiolopolis (<http://radiolopolis.com/>), Docadoc (<http://docadoc.com/>), and Carenity (<http://www.carenity.com/>). These different kinds of social networking sites offer to their members an opportunity to be connected with others and share experiences and knowledge.

Our first goal is to design a social network dedicated to physicians and patients. The basic model of the targeted social media should take into account the management of a:

- Set of patients which provide personal information in their health care profile.
- Set of physicians providing information enabling their identifications.
- Set of mechanisms permitting to patients to upload the medical images related to their diseases.
- Set of mechanisms permitting to physicians to comment the uploaded medical images.
- Set of search functions by which patients and physicians can locate easy and efficient information about medical images.
- Site operator who controls the site and triggers a set of mechanisms permitting to collect medical images in order to process them for various purposes such as medical images’ indexation.

⁷<http://www.asklepios.com/>

⁸<http://www.acc.org/>

⁹<http://www.diabspac.com/>

¹⁰<http://www.parlonscancer.ca/>

¹¹<http://www.renaloo.com/>

¹²<http://www.rxspace.com/>

Like in [7], our social network, addressed to physicians and patients, can connect millions of voices to:

- Increase the timely dissemination and potential impact of health and safety information;
- Leverage audience networks to facilitate information sharing;
- Personalize and reinforce health messages that can be more easily tailored or targeted to particular audiences;
- Empower people to make safer and healthier decisions;
- Facilitate interactive communication, connection, and public engagement;
- Updates patients about changes in physician's practice;
- UKeeps patients informed about upcoming appointments, tests, immunizations;
- Engages patients in discussions about key health issues;
- Answers patients' medical questions;
- Communicates with family members, other caregivers;
- Grows physician's practice;
- etc.

In our work, we respect that our social network needs to contain all of these features, situated above.

2.2 Multilingual Indexation Approaches

2.2.1 An Overview

Annotation, indexing, and retrieval of image content in the large scale online repositories have become an increasingly active field. Annotation and tagging have been recognized as a very important and essential mechanism to enable the effective organization and sharing of large scale of images information. Therefore, efficient automatic annotation and tagging methods are highly desirable. This interdisciplinary research direction has attracted various attentions and resulted in many algorithmic and methodological developments. In the literature, different works of extracting terms from textual corpus use two approaches: statistical analysis and the linguistic or structural analysis. Statistical analysis is based on the study of the contexts of use and distributions of terms in documents. Linguistic analysis uses of language knowledge, such as morphological and syntactic structure of terms. Other works combine the two approaches and constitute an approach called "hybrid or mixed approach."

Several studies [8–11] have used Semantic Resources (RS) in the process of indexing. They find that improving information retrieval systems is based on such resources for indexing multilingual corpus or for detecting lemmatization. They use UMLS (Unified Medical Language System) to index ImageCLEFmed collection. Similarly, thesaurus MeSH (Medical Subject Heading) is used in [11] to index documents of TREC. The same steps are usually used; Starting by language

identification, subsequently the extraction of terms by a linguistic method adapted to the language, finally the projection of extracted terms on semantic resource to detect concepts.

Maisonnasse et al. [8] use a linguistic method of concepts' detection, performed in the collection of ImageCLEFmed 2007. This method was used with three different linguistic tools: MetaMap¹³ is a morphosyntactic analyzer associated with UMLS which can extract concepts from documents. This analyzer deals only with written in English. MiniPar¹⁴ can extract terms in English also. TreeTagger¹⁵ is an analyzer which detects the grammatical category of a word and its lemma. There is a version of TreeTagger for English, French, and German. Although a large number of image tags can be generated, also, in short time, these approaches depend on the availability of human annotators and are far from being automatic. Similarly, research in the other direction via text-to-image synthesis [12–14] has also helped to harvest images, mostly for concrete words, by refining image search engines.

The indexing method used in these works consists of the same steps: identification of the language of the document, extracting terms of this language by a linguistic method adapted to the language, detection of concepts by projecting the extracted terms on semantics resource. Unlike these methods, we would like to show that with an external semantic resource of sufficient quality, it is not necessary to use linguistic tools suited to a given language and a method of extracting purely statistical terms allows obtaining results of equivalent quality. Thus, with this statistical method, we will not have to change language analyzer whenever the document language changes.

2.2.2 Indexation Approaches via Social Networks

The idea behind an “ideal” annotation is to provide recommendations of annotations based on collaborations between users. Several online systems have sprung into existence to achieve annotation of real-world images through human collaborative efforts (Flickr) and stimulating competition [15, 16], in the context of ESP game project. From the same point of view, Google sets up Google Labeler Image a game which consists in determining the common directions of the images for two distant players. This kind of work neglects that the common directions between the users is, in our opinion, strongly related to a social aspect and a preliminary knowledge. The number of images on Facebook, as an example, has exceeded 60 billion by the end of 2010, and around 138 MB of new content is being uploaded every minute. This user-uploaded and user-generated audio-visual content belongs to the established concept of user-generated content (UGC) [17]. UGC includes all kinds of data that come from regular people who voluntarily contribute with data, information, or

¹³<https://metamap.nlm.nih.gov/>

¹⁴http://ai.stanford.edu/~rion/parsing/minipar_viz.html

¹⁵<http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

media that then appears before others in a useful or entertaining way. Automatic tagging and search for image content has been a tremendous challenge, particularly in uncontrolled environments such as UGC applications. Collaborative annotation and tagging has been a typical and promising approach for tagging of user-generated multimedia content [17]. A limited number of studies using data from a social network to improve the suggested annotations are highly recommended. For the purpose of this study, it is essential to integrate or benefit from the use of the following works such as Shevade et al. [18], Stone et al. [19], Messaoudi et al. [5], and Bouslimi et al. [20].

Most approaches to automatic image annotation have focused on the generation of image labels using annotation models trained with image features and human annotated keywords [21–24]. Instead of predicting specific words, these methods generally target the generation of semantic classes (e.g., vegetation, animal, building, places, etc.), which they can achieve with a reasonable amount of success. Recent work has also considered the generation of labels for real-world images. In the same context, Shevade et al. [18] combine measurements of similarity between users. These measurements understand proximities between users in the social network, semantic similarities between concepts using ConceptNet, and of the similarities between events. The annotations are generated while using (1) annotations of the most similar user in the social network, (2) the similarity of the images based on their contents, and (3) the application of the activation spreading on the graph the most similar concepts. The activation spreading is a process of research initiated by the labeling of a set of nodes source with weights. The spreading with the other nodes is iterative, and it takes into account the relation between the nodes.

Zunjarwad et al. [25] have taken the work of Shevade et al. [18]. The contribution of Zunjarward et al. is summarized in what they call social network of confidence. This approach is based on two measures. The first one is founded on a binary value attributed manually by the user to each member of his social network. This binary value corresponds to a value of confidence. However, the second measurement is based on the co-occurrence of the user with others in the same event. The events are tags attributed by the users. A value of confidence is granted to a user, if he has annotated photos with an event that matches the current event annotation.

Stone et al. [19] proposed a solution to identify people on Facebook using the recognition and similarity of faces by content. To suggest the figurants' user names, they estimate statistically the intensity of inter-user relationship. To proceed this method, one should consider two metrics which are: the number of photos of a user on which a person is identified or the pictures of his friends and that the number of pictures when user is present together with that person.

Manning et al. [26] have proposed a collaborative approach in which patients seek via a social network to find quick analysis of their medical images by expert doctors. In addition, they use a mixed approach terminology extraction annotations. Once the terms found, they seek in MeSH thesaurus the concepts that are related to the keywords already found.

Fuming et al. [27] have also proposed a collaborative approach, wherein several different statistical models are effectively combined to predict the annotation for

each image. In addition, they combined the correspondence between keywords tokens and visual image/regions and word-for-word correlation to improve the annotation.

Sun et al. [28] have combined the annotation of similar images via collaborative approach. Similar images are searched with search engines, and their tags then infer via word correlation the annotation of target image.

Bouslimi et al. [20] have suggested an automatic system for medical image annotation that combines textual and visual descriptors using latent semantic vector to build a semantically medical image. Indeed, to automatically annotate a new medical image, they compare the vector describing the source image with the vectors that are already existing in the database.

Messaoudi et al. [5] have proposed a collaborative approach, in which patients seek via a social network to find quick analysis of their medical images by expert doctors. In addition, they use a mixed approach terminology extraction annotations. Once the terms found, they seek in MeSH thesaurus, the concepts that are related to the keywords already found.

Kanishcheva and Angelova [29] have suggested an approach to image auto-tagging refinement. They have presented a post-processing tool that refines tags associated with images. The tool works stepwise in two phases. There are several steps in phase 1 that maintain tags as English wordforms. Step 1.1 cleans mistakes (as tags sometimes contain typos) and splits words which are written together. Step 1.2 finds plural forms and transforms them to singular; it also analyzes the inflection. Step 1.3 analyzes synonyms. Step 1.4 analyzes phrases. Step 2, about semantic analysis that consolidates tags using English language resources like WordNet, will be considered below.

Bouslimi and Akaichi [30] have proposed a social network for collaboration whose goal is learning among medical residents that are currently in the course of their training in radiology. Indeed, the terminology extraction from comments has allowed them to obtain the relevant terms and using a correlation with the MeSH thesaurus in order to have access to get the concepts. They used the keyword results for the construction of an annotation of the medical image.

We have made a comprehensive review on the state of the art of medical images' indexation. The automatic extraction tools help terminologists to validate the extracted terms. We note the existence of a wide variety of extraction tools. Linguistic techniques have a fine linguistic description and have the ability to handle small corpus. However, it requires a large linguistic knowledge. The statistical approaches have the advantage of requiring no prior knowledge of language and are applicable on corpus for which no external resource (dictionary, stop list, ontology, etc.) has been developed. The results of statistical approaches are strongly related to the corpus studied and cannot be generalized beyond this context. These approaches are efficient for corpus sufficiently large size. They are not applicable on small corpus sizes. Finally, the hybrid approaches, which provide quality results, present a compromise between statistical methods and linguistic methods. The idea of combining these two methods is relevant. Indeed, this combination takes advantage of the fine linguistic analysis and robustness digital analysis. The hybrid approach

takes advantage of the speed and independence from the field of statistical methods. Indeed, Harrathi [31] demonstrated that the use of methods of extracting purely statistical terms provides results of equivalent quality. For author, it is not necessary to use linguistic tools adapted to a given language. He shows that with an external semantic resource of sufficient quality, statistical approach gives greater results than those using linguistic techniques. In addition, the statistical approach is simple to implement as opposed to linguistic approaches. Our approach is similar to that of Manning et al. [26], and inspired from the work of Bouslimi and Akaichi [30] and Messaoudi et al. [5], through which we propose a social network for the collaboration between physicians and patients. But, most of these works do not cover all languages. They present a monolingual terminology extraction based on the extraction of terms in a specific language: the French language or the English language. We will present, in our work, a medical images multilingual indexation from comments, based on statistical methods and on external semantic resources. Our approach allowed us to obtain the relevant multilingual terms. After that, we get the multilingual concepts by using the correlation with the multilingual MeSH thesaurus. In addition, the multilingual keywords results construct an annotation of the medical image, in different languages. In what follows, we explicitly describe the description of our social network, firstly and our indexation's approach, secondly. We added a correction of spelling mistakes using the medical vocabulary to improve our indexing process, inspired from the work of Bouslimi and Akaichi [30].

3 Social Network Architecture Description and Implementation

Social network is the study of social entities, as a group of people linked to one another by one or more common attributes, and their interactions and relationships. The interactions and relationships can be represented with a network or graph, where each node represents an actor and each link represents a relationship. A social network plays an important role as a support for the dissemination of information, ideas, and impact among its members. It may be a major selling point for patients looking for physicians. In fact, social networking has been proven effective in sharing knowledge and establishing communication among patients and physicians.

Our proposed social network, as many others described in the literature Gong and Sun [32] and Almansoori et al. [33], aims to allow patients and physicians to connect with each other by eliminating all geographical and time frontiers. Both users exploit it to seek advices and to share experiences related to medical images' interpretations and diseases' analysis.

Like in Facebook, patients and physicians need to be registered by creating profiles about themselves (detailed information). This step is very important in order to use our medical social network. Our social network site is a virtual place where registered patients upload their medical images to be commented

by various registered physicians. Not only registered patients can upload their medical images, but even registered physicians have this right, especially medical students, in the objective to learn from this collaboration. Uploading and sharing images are not the only functionalities. Users can also sharing status, medical events, and any information related to medical activities. The main objective, behind using our medical social network, is to enable patients and physicians to exchange information, share knowledge, and experiences. It, also, gives to its members the opportunity to connect and communicate with others in need of support and encouragement related to different situations and health care problems. Consequently, the patient's condition represents the heart of the health system justified by experts' interpretations and analysis expressed by comments. Patients frequently place trust in peer recommendations on social network site making them key platforms to influence change. We used PHP (Hypertext Preprocessor), in the development of our social network, and MySQL for the database management. PHP is a server-side scripting language not only designed for web development but also used as a general-purpose programming language. It ensures many features that allow creating and managing an entire social network website. Pages were written in HTML, PHP, Ajax, and JavaScript and they were designed using Dreamweaver. In order to better explain our medical social functionalities, we describe in the following some interfaces of the social network:

- The first interface to use is the identification interface. Patients and physicians need to provide the registration process, as new users. The registration interface dedicated to physicians has more specificity because profiles have different structures.
- After the identification, the basic functions and services of the social network are displayed in the main interface. Patients and physicians can perform different activities, especially, get access to posting, commenting functions, and uploading images, etc.
- Mixed posts and comments, in different languages, performed by different patients and physicians are entered and displayed, in the posting interface. Posts will be treated by back office functions implementing our multilingual indexation's approach.
- To answer urgent questions and/or advices, synchronous communications between members (patients and physicians, physicians and physicians) are performed in the communication interface. Mixed posts performed by different physicians and patients, according to their opinions and questions, are displayed in the posting interface.

The different interfaces of our social network allow an easier mapping for users, their intent, and targeted functions. The most important is that these interfaces describe the way of interaction between patients and physicians with the social network site and the way of interaction between patients and physicians with each other by exploiting existing functions. The social network content blocks are visually separated. This separation will help to organize pages content and each element is well defined and presented separately, to succeed the interaction between

patients and physicians. In fact, this separation makes the content understandable, easy to recognize and makes the content more reachable. In order to improve the interaction between users (patients and physicians, physicians and physicians), our social network contains an advanced search function, allowing the organization of the connections between patients and physicians having common interests related to diseases and their analysis, interpretations and treatments. This function supports users to rapidly find the content and contacts they are searching for. Considering the recommendations of a design expert, the created interfaces are simple in terms of color scheme and graphics, for example, in Facebook. The idea consists of using a few colors and the background is generally white. This management of color scheme helps the exploitation of our social network site by physicians and patients, makes the content well presented, and comments better seen and readable. Therefore, buttons and links are placed almost on every page of the social network. Some links are related to the navigation processes and some others permit to users to regulate specific functions. Buttons are used to associate users to actions and to navigate among different pages; they are clear and more remarkable.

When incoming comments or messages appear on the social network sites, physicians and patients need to react to them in real time. For this reason, they should consider establishing their own social network presence. So, we assure our social network with a synchronous communication, in order to establishing interaction and sharing information. We, also, equipped our social network with a real-time update feature ensuring the delivery of updates (medical images uploading, physicians' annotations, etc.) as soon as they are submitted. In order to encourage patients and physicians to exploit our social network site, many actions could be performed such as suggesting new friends, preferably in relation with medical activities, interests, events, and groups. This is performed to extend their social circles allowing implicitly the extension of the shared knowledge about medical images and the associated diseases.

The extraction of more knowledge, from various posts or comments, is possible by using mining tools. For this reason, our social network is equipped by mining tools such as those presented in Xie et al. [34], where authors consider the emergence and pervasively of online social networks that have enhanced web data with developing interactions and communities both at large scale and in real-time.

The successful building of our social network needs to respect security aspects and to protect users' privacy. Physicians must protect their own privacy, as well as that of their patients. To do it, we use privacy settings on our social network site to keep out everyone except patients or fellow physicians. We will also consider security aspects such as those presented in [35], where the author proposed methods to secure healthcare social networking sites providing users with tools and services to easily establish contact with each other around shared problems and utilize the wisdom of masses to outbreak disease.

The successful building of the social network site needs also to respect a privacy policy considered to protect the users whether they are patients or physicians. The American Medical Association [36] recently adopted a policy to ensure profes-

sionalism and protect patients' privacy during social network activities. According to the AMA policy (AMA 2012), physicians should:

- Never post identifiable patient information;
- Separate personal from professional content;
- Use privacy settings when going on social networking sites, to protect personal information;
- Monitor their sites and their presence on other Internet sites, to make sure content is appropriate and accurate.

The AMA policy also warns that anything doctors post on the Internet may become public and permanent. The AMA [36] explains inappropriate postings can have a detrimental effect on a doctor's reputation among patients and colleagues. Because such postings can also be harmful to the medical profession, the AMA advises physicians also to be alert for unprofessional content placed on the Internet by other physicians. And if doctors see something, they are obligated to do something about it (AMA 2012). In the same time, physicians must protect their own privacy, as well as that of their patients. Once doctors have a presence on the Internet, though, privacy may be difficult to maintain. They must exercise caution about the information they put on social network sites—both their own and others. Moreover, it is essential that physicians avoid sites that could prove awkward if their visit there were revealed [36].

4 The Proposed Methodology

Figure 2.1 shows a structure of the conceptual design for our multilingual approach which we consider as a mixed method involving the combination of statistical methods with a multilingual external semantic resource. From Fig. 2.1, we can see that our methodology mechanism has four main components: pre-processing unit; cleaning, correcting, and lemmatization unit; terms' extraction unit, and concepts' detection unit.

Each of these steps, as shown in Fig. 2.1, consists of several separate processes: cleaning, lemmatization, multilingual simple terms' extraction, multilingual compound terms' extraction, and multilingual concepts' extraction. Our method of multilingual concept's detection is a preliminary step to a semantic indexing process. Our terminology extraction approach is a mixed approach which combines the two approaches linguistically and statistically. The algorithm, describing the steps of the proposed method in generally, is below.

As described by this algorithm, the multilingual indexation's approach initiates by the creation of the textual corpus containing comments extracted from the medical social network and performed on a medical image. Then, we will describe the different steps of our approach step by step.

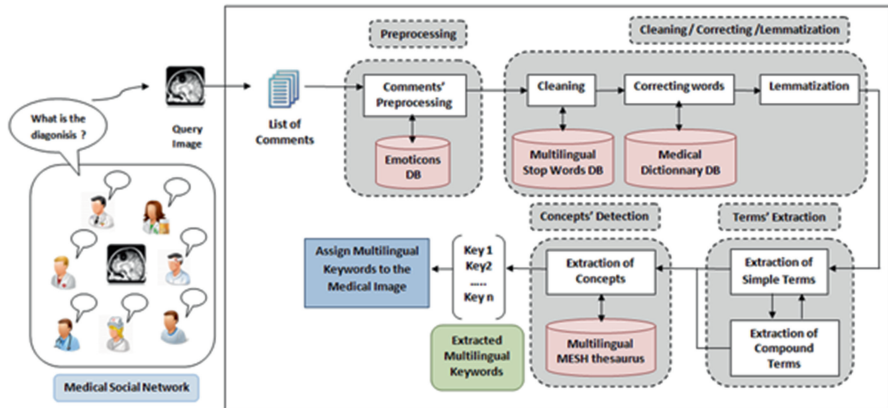


Fig. 2.1 Proposed methodology pipeline

MULTILINGUAL INDEXATION'S APPROACH ALGORITHM

INPUT
 C_i : A Comment in the Corpus of comments $\{Comment\}$
OUTPUT
 MC_{on} : Multilingual Concepts $\{List\ of\ full\ Multilingual\ Keywords\}$
VAR
 MS_t : Multilingual Simple Terms

 MC_t : Multilingual Compound Terms

 M_{esh} : Multilingual MeSH Thesaurus

 L_w : A List of Multilingual Words
0. BEGIN

1. **PREPROCESS** OF C_i $\{Comment\}$ **IN** List of Comments
2. $L_w \leftarrow$ **CLEAN** C_i $\{Comment\}$ **USING** a Multilingual Anti-Dictionary;
3. **FOR EACH** W_i $\{word\}$ **IN** L_w **DO**
4. **CORRECTING** (W_i) **USING** a Multilingual Medical Dictionary;
5. **PERFORM** Lemmatization (W_i);
6. **END FOR**;
7. $MS_t \leftarrow$ **EXTRACT** Multilingual Simple Terms (L_w);
8. $MC_t \leftarrow$ **EXTRACT** Multilingual Compound Terms (MS_t , L_w);
9. $MC_{on} \leftarrow$ **EXTRACT** Multilingual Concepts (MS_t , MC_t , M_{esh});
10. **END**.

4.1 Comments' Pre-processing

The first step consists of the pre-processing operation that removes punctuation such as (,;?;!,[],)Etc..) and cuts the sentence into words. We remove also the emoticons' symbols, by using an anti-dictionary that contains the emotions that are used frequently in twitter, Facebook, etc. Our pre-treatment algorithm contains five procedures, one for each sub-step. The algorithm will be presented below contains a procedure to replace the emoticon's symbol and the punctuation space by character

space, then a procedure that convert text to lowercase, procedure that Clean Text and finally procedure Split text into words. So, the pre-processing step is composed of five preliminary stages:

- Decomposition of the corpus sentences.
- Removing the punctuation points and cleaning sentence.
- Removing the emoticons' symbols.
- Converting sentence to lowercase.
- Cutting the sentence into words.

The algorithm, describing the pre-processing phase of the proposed method, is below.

PRE-PROCESSING ALGORITHM

INPUT

C_i : A Comment in the Corpus of comments {*Comment*}

OUTPUT

L_{word} : List of Words

VAR

S_m : A Sentence in the list of sentences

Low_s : Lowercased Sentence

0. BEGIN

1. **FOR EACH** C_i {*Comment*} **IN** List of Commentary **DO**
2. **DECOMPOSITION OF** C_i {*Comment*} **INTO** Sentences;
3. **FOR EACH** S_m {*Sentence*} **IN** Sentences **DO**
4. **DELETE** the emoticons' symbols **SINCE** emoticons' DB;
5. **REMOVING** the punctuation points;
6. **CLEANING** (S_m);
7. $Low_s \leftarrow$ **LOWERCASE** (S_m);
8. $L_{word} \leftarrow$ **CUTTING** (Low_s) **INTO** words;
9. **END FOR**;
10. **END FOR**;
11. **END.**

4.2 Cleaning, Correcting, and Lemmatization

4.2.1 Cleaning

The cleaning step is about removing the stop words. A stop word (empty word) is a word which is not useful to index a document such as prepositions, articles, pronouns, some adverbs and adjectives, and finally some frequent word. Each language possesses a list of stop words which corresponds to his specificity. This type of word has a low informative power. To assure this task, we use a multilingual anti-dictionary containing the blackest words, in different languages, that seem useless in the medical field. We cite among them, from three different languages:

(many, how, again, which, since then, this, on some, but there, like why, however, when, which, soon, etc.) in English; (alors, au, aucuns, aussi, autre, avant, avec, avoir, bon, car, ce, etc.) in French, and (aber, als, am, an, auch, auf, aus, bei, bin, bis, bist, da) in German. The rest of the words (full words) will be used to describe the images contents because it has a high discriminatory power.

4.2.2 Correcting Words

Comments may include many misspelled words. These misspelled words may have a negative effect on our indexation’s approach and on the search function later. For this reason, correcting words is based on using of a medical dictionary to detect misspelled words. A word is considered misspelled when it does not appear in the dictionary. In this case, we presented to use the LEXILOGOS¹⁶ Dictionary of medical terms. LEXILOGOS can cover many languages especially from the European countries. There are many mathematical metrics providing a measure of similarity between two character sequences. According to Bouslimi and Akaichi [30], the use of the two distances Levenshtein [37] and Jaccard [38], to correct spelling mistakes and to find the nearest word to be corrected, is highly efficient. The performance of the two distances has been proven, also, by Heasoo et al. [39]. According to Navarro [40], Levenshtein distance can be used to generate possible corrections. This distance is equal to the number of characters one has to delete, insert, or replace to go from one sequence to another. In most cases, the distance between a misspelled word and its correction is 1. All candidate corrections will be made up of all words obtained by:

- Inserting a character in the misspelled word;
- Deleting a character of the misspelled word;
- Substituting a character of the misspelled word.

The formula expressing the Levenshtein distance is the following:

$$\begin{aligned} \max(i,j) & \qquad \qquad \qquad \text{if } \min(i,j) = 0, \\ lev_{a,b}(i,j) = \min & \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{a_i \neq b_j} \end{cases} & \text{otherwise} \end{aligned} \tag{2.1}$$

We will present, thereafter, two examples of operations they need to correct English and French words.

Example 1 (Lev hemoragiie, Hémorragie)

h	é	m	o	r	r	a	g	i		e
	replace				insert			replace	delete	
h	e	m	o	r		a	g	i	i	e

¹⁶http://www.lexilogos.com/medical_dictionnaire.htm

Example 2 (Lev vontrycular, Ventricular)

v	e	n	t	r	i	c	u	l	a	r
	replace				replace					
v	o	n	t	r	y	c	u	l	a	r

Like in [30] and for more precision, we used the Jaccard distance to support the Levenshtein distance. The Jaccard distance is the ratio between the cardinality (size) of the intersection of the sets considered and the cardinality of the union of the sets. It evaluates the similarity between the sets. The words w_1 and w_2 are represented, not as vectors, but as letter sets. The similarity obtained is $d_{jaccard}(w_1, w_2) \in [0, 1]$. The formula expressing the Jaccard distance is the following:

$$d_{jaccard}(w_1, w_2) = \frac{\vec{w}_1 \cdot \vec{w}_2}{|\vec{w}_1| |\vec{w}_2| - \vec{w}_1 \cdot \vec{w}_2} \quad (2.2)$$

The choice of the most probable correction is done by attributing to each correction candidate a score. The higher the score is, the more likely it is. That is to say, the candidate correction is the correct spelling of the word to correct [30].

4.2.3 Lemmatization Words

In other side, comments may include different forms of words belonging to various families, for grammatical reasons. Lemmatization, that seeks the canonical form of words, is the process of grouping together the different inflected forms of a word so they can be analyzed as a single item: the name, the plural, the verb in the infinitive, etc. It is used to regroup words in their belonging family. To find the lemmas, we implement the stemmer algorithm which seeks the root (prefix) and then assigns the suffix for a parent noun. According to Frakes and Fox [41], stemming algorithms are used in many applications related to natural language processing such as text analysis systems, information retrieval, and database search systems. We update and use Porter algorithm [26] which is considered as the best known manner corresponding to our needs. Table 2.1 presents an example of rules for English language from different steps of the Porter stemmer.

Table 2.1 An example of rules

Example of rules	Example of results
Rule1: sses \rightarrow ss	Caresses \rightarrow caress
Rule2: ator \rightarrow ate	Operator \rightarrow operate
Rule3: ness \rightarrow to remove	Goodness \rightarrow good
Rule4: ible \rightarrow to remove	Defensible \rightarrow defens

We implemented the entire algorithm using all grammatical rules associated with French, English, German, Spanish, and Italian languages, in order to validate our method. The algorithm, describing the Cleaning, Correcting, and Lemmatization phase of the proposed method, is below.

4.3 Terms' Extraction

We distinguish two types of terms: simple terms composed of a single word and the compound terms composed of a sequence of words. We used a multilingual anti-dictionary, containing multilingual stop words, for extracting simple terms and a statistical measure for the extraction of compound terms.

CLEANING CORRECTING AND LEMMATIZATION ALGORITHM

INPUT

W_i : A Word in the List of Words

OUTPUT

L_{ifw} : A List of Lemmatized Corrected Words

VAR

ST_w : A Stop Word in the Stop Words DB

W_{md} : A Word in the Medical Dictionary

d_1 : Levenshtein Distance Value

d_2 : Jaccard Distance Value

0. BEGIN

```

1. FOR EACH  $W_i$  {Word} IN List of Words DO
2.   FOR EACH  $ST_w$  {Stop Word} IN Stop Words DB DO
3.     IF  $W_i$  IS A  $ST_w$  THEN
4.       PERFORM CLEANING ( $W_i$ );
5.     ELSE IF  $W_i$  IS NOT A  $ST_w$  THEN
6.       FOR EACH  $W_{md}$  IN Medical Dictionary DO
7.          $d_1 \leftarrow \text{Calculat\_Levenshtein}(W_i, W_{md})$ ;
8.          $d_2 \leftarrow \text{Calculat\_Jaccard}(W_i, W_{md})$ ;
9.         CORRECTING ( $W_i, W_{md}, \text{MAX}(d_1, d_2)$ );
10.         $L_{ifw} \leftarrow \text{LEMMATIZE}(W_i)$ ;
11.      END IF;
12.    END FOR;
13.  END FOR;
14. END.
```

4.3.1 Simple Terms' Extraction

According to Bouslimi and Akaichi [30], many approaches seek to define a key term based on some statistical features and studying their relation with the notion

of importance of a candidate term. If a candidate is considered the important term in a document analyzed, then it is relevant as a key term. TF-IDF of Jones [42] and Likey of Paukkeri and Honkela [43] are two methods which compare the behavior of a candidate term in the analyzed document with his behavior in a collection of documents (reference corpus). The objective is to find candidate terms whose behavior in the document that varies positively compared to their global behavior in the collection. The two methods were expressed by the fact that a term has a strong importance face to face of the analyzed document if there is present, then it is not in the rest of the collection. According to the same study, TF-IDF gives good results compared to Likey, and that is why in the extraction step of the keywords we have used this method. This method combines two factors, the local weighting (TF) which quantifies the local representation of a term in the corpus of comments and the overall weighting (IDF) which measures the global representative of the term on the collection of corpus based on provided comments. We will keep only the terms that the output value exceeds, according to a threshold fixed to 0.125. The formula expressing the measure is the following:

$$TF - IDF(term) = TF(term) * \log \frac{N}{DF(term)} \quad (2.3)$$

where TF represents the number of occurrences of a term in the analyzed document, DF represents the number of documents in which it is present, and N is the total number of documents. The higher the score of TF-IDF of a candidate term, the more it is important in the analyzed document. The algorithm, describing the simple terms' extraction step of the proposed method, is as follows:

SIMPLE TERMS' EXTRACTION ALGORITHM

INPUT

W_i : A World in the List of Lemmatized Words

C_i : A Comment in the Corpus of comments

OUTPUT

L_{st} : A List of Simple Terms

CONSTANT

$T_{hrd} = 0.125$ // The Threshold

VAR

W_{eit} : float // The Weighting of The Word in The Comment

0. BEGIN

1. **FOR EACH** C_i {Comment} **IN** List of Commentary **DO**
2. **FOR EACH** W_i {World} **IN** List of Lemmatized Words **DO**
 // Calculate The Weighting of The Word
3. $W_{eit} \leftarrow TF(term) * \log \left(\frac{N}{DF(term)} \right)$
4. **IF** $W_{eit} > T_{hrd}$ **THEN**
5. $L_{st} \leftarrow W_i$;
6. **END IF**;
7. **END FOR**;
8. **END FOR**;
9. **END.**

4.3.2 Compound Terms' Extraction

This step is about the identification if the term is a compound or a single word. A collocation is a combination of recurrent words which are more often found together than each one of them alone. Some collocations are fixed noun phrases that are specific to a domain [31]. Mutual Information, as a measure, is used to extract relevant collocations. This measure compares the probability of co-occurrence of words and the probability of each words separately [5]. The process of extracting complex terms is iterative and incremental. We build complex terms of length $n + 1$ words from the words of length n . For each sequence of words, we compute the value of the MI . If the sequence of words exceeds the threshold set to 0.15 in our case, the sequence will be comprised on the list of compound terms. Note that the computation of MI is ensured by the following formula:

$$MI(m_1, m_2) = \log_2 \frac{P(m_1, m_2)}{P(m_1) * P(m_2)} \quad (2.4)$$

where $P(m_1)$ and $P(m_2)$ are an estimation of the probability of occurrence of the words m_1 , m_2 and $P(m_1, m_2)$ is an estimation of the probability that the two words appear together. The algorithm, describing the compound terms' extraction step of the proposed method, is as follows:

COMPOUND TERMS' EXTRACTION ALGORITHM

INPUT

L_{st} : A List of Simple Terms

L_{word} : List of Words

OUTPUT

L_{ct} : A List of Compound Terms

CONSTANT

$T_{hrd} = 0.15$ // The Threshold

VAR

MI : The Mutual Information

W_i : A World in The Corpus

t_i : A Term in The List of Simple Terms

0. BEGIN

1. **FOR EACH** t_i {term} **IN** L_{st} **DO**
2. **FOR EACH** W_i {World in The Corpus} **IN** L_{word} **DO**
3. $MI \leftarrow \log_2 \frac{P(t_i, W_i)}{P(t_i) * P(W_i)}$
4. **IF** $MI > T_{hrd}$ **THEN**
5. $L_{ct} \leftarrow (\text{CONCAT}(t_i, W_i));$
6. **END IF;**
7. **END FOR;**
8. **END FOR;**
9. **END.**

4.3.3 Concepts' Extraction

After extracting simple and compound terms which participate to candidate keywords, it comes to the extraction of concepts which are chosen from a controlled vocabulary (a dictionary, thesaurus, or a list of terms, etc.). This is a verification step which comes to use an external semantic resource. We use, in our case, a multilingual Mesh thesaurus to cover different languages. We extract concepts by projecting those terms on the thesaurus. More precisely a medical multilingual thesaurus MeSH is used to filter keywords obtained in the previous step, in order to verify that the extracted term belongs, or not, to the medical vocabulary. We used the UMLS (Unified Medical Language System) provided by the National Library of Medicine. It is a multilingual meta-thesaurus covering the medical field. This resource was created to facilitate research and the integration of information from multiple electronic sources of biomedical information [44]. It is the fusion of several semantic resources. UMLS contains more than one million of concepts related to more than 5.5 million terms in 17 languages. The 17 languages are not covered in the same way in UMLS. English is the language most represented with 68% of the vocabulary, German vocabulary covers 2.84%, and French vocabulary covers 2.55%. The algorithm, describing the concepts' extraction step of the proposed method, is as follows:

CONCEPTS' EXTRACTION ALGORITHM

INPUT

L_{st} : A List of Simple Terms
 L_{ct} : A List of Compound Terms

OUTPUT

L_{con} : A List of Concepts

VAR

S_{ti} : A Term in The List of Simple Terms
 C_{ti} : A Term in The List of Compound Terms

0. BEGIN

1. **FOR EACH** S_{ti} {simple term} **IN** L_{st} **DO**
2. **IF** S_{ti} **IN** {Thesaurus MeSH} **THEN**
3. $L_{con} \leftarrow S_{ti}$;
4. **END IF**;
5. **END FOR**;
6. **FOR EACH** C_{ti} {compound term} **IN** L_{ct} **DO**
7. **IF** S_{ti} **IN** {Thesaurus MeSH} **THEN**
8. $L_{con} \leftarrow C_{ti}$;
9. **END IF**;
10. **END FOR**;
11. **END.**

5 Experimental Results

After developing the medical social network site, the challenge is to extract relevant multilingual terms and keywords from comments to index and annotate images.

5.1 Data Test and Evaluation Criteria

Experiments have been carried out to validate the efficiency of the proposed model. The experiments were carried out on a Core i_3 , 2.4 GHz processor with 4 GB RAM using NetBeans editor. The Multilingual Anti-Dictionary is saved in an XML file. The files will be examined using JDOM (Java). For querying the MESH¹⁷ thesaurus (meshdata.rdf) file, we used the java API Sesame of Thomas Francart which allows executing applications written in SPARQL queries.

The relevance of our approach was evaluated on preselected medical images, which are supplied by internal physicians and residents during their training, diagnostic, and images analysis in relation with patients' states and clinical cases, at Charles Nicolle Hospital in Tunis. This collection contains 200 medical images annotated by 40 physicians, involved in various specialties, from different countries. Different analyses are the results of the collaboration between Tunisian physicians and other physicians from countries like France, UK, and Germany. We collected 100 examination cases in different languages, where each examination consists on a comment related to a medical image. Analyses are essentially written in French, English, Italian, Spanish, and German.

We also perform experiments on the collection of ImageCLEF'2013. This collection is composed of over 45,000 biomedical research articles in PubMed Central (R). Each document is constituted of a medical image and a portion of text. The collection contained over 300,000 images including MR CT, PET, ultrasound, and combined modalities in one image. The images are very heterogeneous in size and content.

We evaluate the performance of this approach by using three measures frequently used in the evaluation of information retrieval systems, namely the precision, the recall, and the MAP (Mean Average Precision) metrics. The MAP represents the quality of a system based on different levels of precision [26]. The precision is the number of correct concepts divided by the total number of extracted concepts. The equation of the precision is

$$\text{Precision} = \frac{\text{The number of correct concepts}}{\text{The total number of extracted concepts}} \quad (2.5)$$

¹⁷<http://www.nzdl.org/Kea/download.html>

The recall is the ratio between the correct terms and the total number of correct concepts that should have been extracted. The equation of the recall is

$$Recall = \frac{\text{The number of correct concepts}}{\text{The total number of correct concepts that should have been extracted}} \tag{2.6}$$

5.2 Evaluation and Results of Our Approach

Our study has been started with a pre-processing phase of the annotations where we have used our own algorithm that cleans the stop words and emoticons' symbols in the text. We used the indexation's method to lemmatize comments and to extract relevant medical terms intended to be used for medical images annotation and indexation. A medical dictionary is used that contains all vocabularies used in medicine to correct errors found during the indexation. This correction is based on the use of the combination of the two distances of Levenshtein and Jaccard to evaluate the correlation between the correct term and the terms found in the medical dictionary. We attempt then to prove that results are confirmed in practice. Figure 2.2. presents an example of comments on cranial CT scan, performed by six physicians. Indeed, according to this example, we notice the existence of an ambiguity of using different languages to express diagnostic and analysis and another ambiguity of the existing of incorrect words.

Table 2.2 shows the calculation that has been done during the correction phase, in relation with multilingual terms existing in Fig. 2.2.

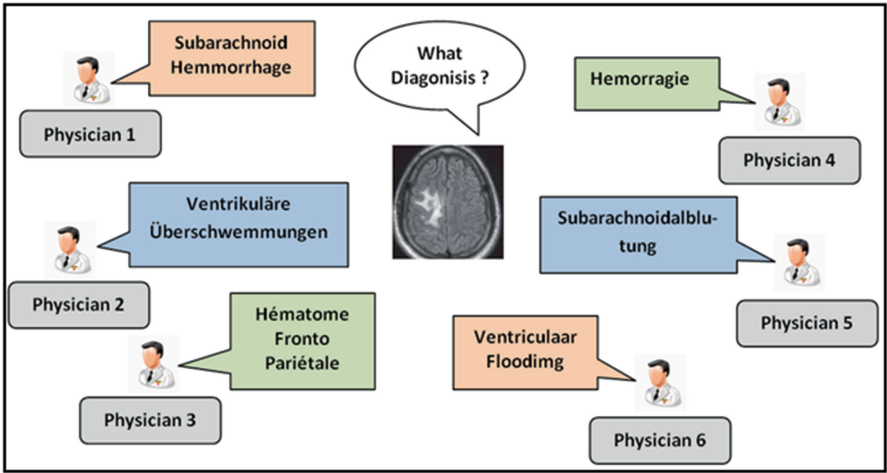


Fig. 2.2 Example of cranial CT scan commented by physicians

Table 2.2 Table of correct words

List of words	Levenshtein distance	Jaccard distance	Max (d_J , d_L)	Correct words
Hémorragie	0.85	0.65	0.85	Hémorragie
Ventrikuläre Überschwemmungen	1	1	1	Ventrikuläre Überschwemmungen
Hématome FrontoPariétale	1	1	1	Hématome Fronto Pariétale
Ventriculaar flooding	0.92	0.88	0.92	Ventricular flooding
Subarachnoidalblutung	1	1	1	Subarachnoidalblutung
Subarachnoid hemmorrhage	1	1	1	Subarachnoid hemmorrhage

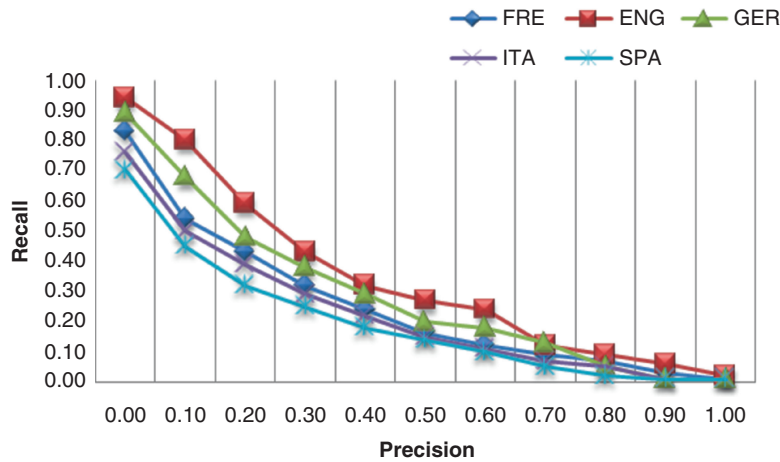


Fig. 2.3 Precision versus recall curves for English, German, French, Italian, and Spanish languages

We perform some experiments to check the effectiveness of the proposed method. The application of the precision-recall curve enables the evaluation of the proposed scheme. It is clear, after many tests, from Fig. 2.3. that proposed method works very well with five different languages: French, English, Italian, Spanish, and German. Figure 2.3 shows the curves of average precision to 11 points of recall with our method of concepts' detection. The curves show that the coverage of the language has a direct impact on system's performance. Indeed, we find that the accuracy obtained for English language is larger than other languages. UMLS covers English better than other languages. The average precision obtained for French and German languages are almost similar with a slight improvement for German language. These two languages are covered in the same manner in UMLS. German language has a slightly higher coverage than French language in UMLS. The average precision obtained for Italian and Spanish languages are almost similar. These two languages are also covered in the same manner in UMLS.

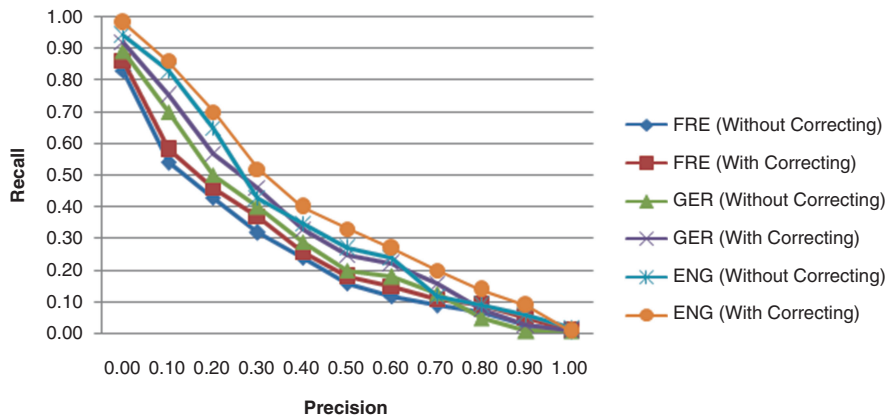


Fig. 2.4 Precision versus recall curves for English, German, and French languages with and without the spelling correction phase

In order to evaluate the performance of our approach, Fig. 2.4. presents a comparative study of our approach firstly without taking into consideration the spelling correction phase and secondly with the spelling correction phase. The spelling correction phase has a positive effect on the indexation process.

We report experimental results that show the feasibility and utility of the proposed algorithm and compare its performance with state-of-the-art methods: we conducted a comparative study with three linguistics methods presented by Maisonnasse et al. [8]. The comparison is carried out by an application of each method and record different results from tests, performed on our dataset. We conducted, also, a comparative study with the works of Messaoudi et al. [5] and Bouslimi and Akaichi [30]. Authors, in these works, handle the same dataset. It allows us to compare our results with published results. Figure 2.5 shows that our proposed system performance is better than three linguistics methods presented by Maisonnasse et al. [8]. We take into account the comparison of our approach with and without the spelling correction phase.

Figure 2.6. shows that our proposed system performance is also better than the other systems presented by Messaoudi et al. [5] and Bouslimi and Akaichi [30].

Table 2.3 presents the performance of each system using the hybrid approach in its indexing and we record results from 299,764 words annotated manually. Figure 2.7 shows the comparison of the proposed system with other systems in terms of the Precision and the Recall. It shows the Precision and the Recall values of each system via different values from Table 2.3 by a vertical bar.

Table 2.4 shows that our proposed system performance is better than all systems in term of Mean Average Precision, after many tests. Figure 2.8 shows the comparison of the proposed system with other systems in term of Mean Average Precision. It shows the Mean Average Precision value of each system via the value from Table 2.4 by a vertical bar.

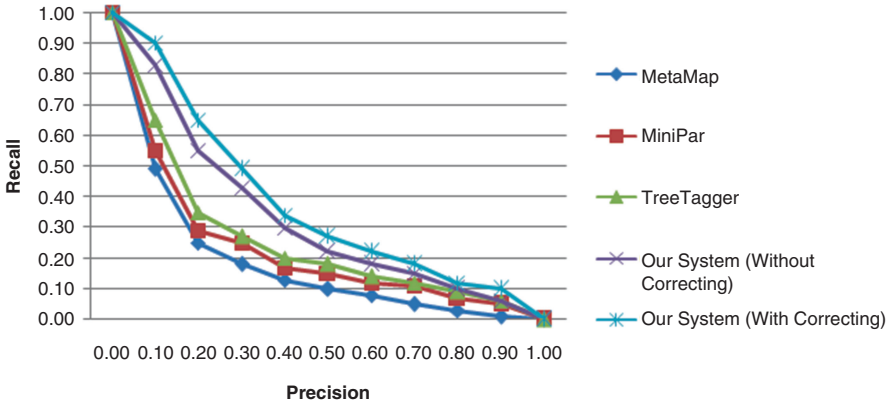


Fig. 2.5 Precision versus recall curves for the comparison between our system (with and without the spelling correction phase) and MetaMap, MiniPar, and TreeTagger systems

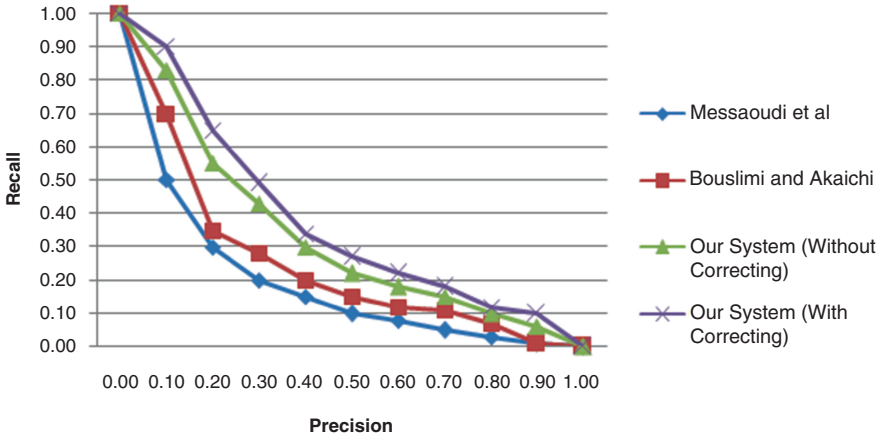


Fig. 2.6 Precision versus recall curves for the comparison between our system (with and without the spelling correction phase) and Messaoudi et al. and Bouslimi and Akaichi systems

For more precision, we evaluated the performance of our solution with that of Messaoudi et al. [5] and Bouslimi and Akaichi [30] by using 2 corpus of CRTT,¹⁸ which are composed by articles taken from the base Science Direct. We see from the figures below, our solution provides good results by the subscribers to the results obtained by Messaoudi et al. [5] and Bouslimi and Akaichi [30].

¹⁸http://perso.univ-lyon2.fr/~maniezf/Corpus/Corpus_medical_FR_CRTT.htm

Table 2.3 Precision and recall of different systems

System	Automatically extract	Corrects	Precision	Recall
MetaMap	263,465	168,631	0.64	0.56
MiniPar	258,752	170,893	0.66	0.57
TreeTagger	264,533	185,269	0.70	0.61
Messaoudi et al. [5]	265,674	191,556	0.72	0.63
Bouslimi and Akaichi [30]	261,181	190,662	0.73	0.63
Our system (without correction)	265,820	207,340	0.78	0.69
Our system (with correction)	265,820	215,315	0.81	0.71

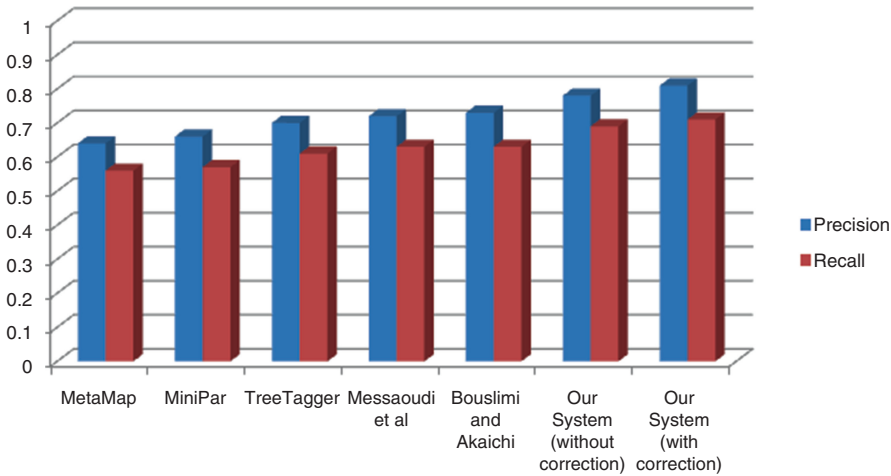


Fig. 2.7 Precision/recall comparisons between the proposed system and existing systems

Table 2.4 Average precision comparison between the proposed system and existing systems

	MAP
MetaMap	0.246
MiniPar	0.246
TreeTagger	0.258
Messaoudi et al. [5]	0.262
Bouslimi and Akaichi [30]	0.269
Our system (without correction)	0.272
Our system (with correction)	0.276

The extracted multilingual keywords and concepts are directly attributed to the medical image. The use of the multilingual thesaurus MeSH to extract medical concepts constitutes a positive point. The selected multilingual terms shall be defined such as keywords to automatically annotate the medical image through the medical social network site (Figs. 2.9 and 2.10).

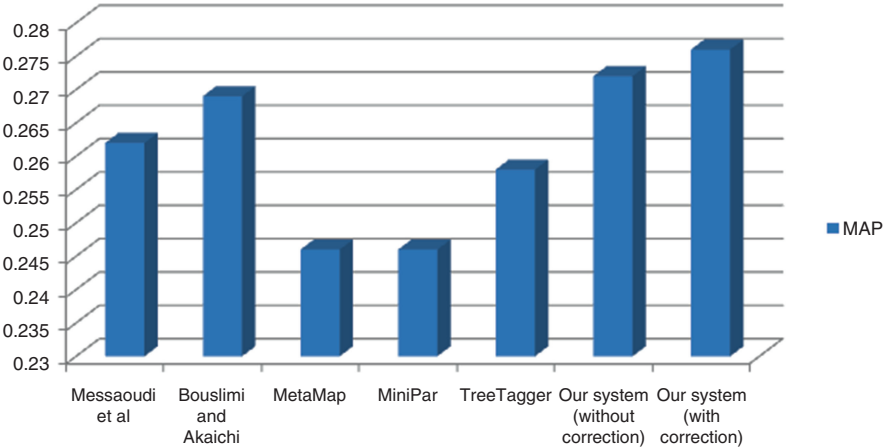


Fig. 2.8 Mean average precision comparisons between the proposed system and existing systems (with and without the spelling correction phase)

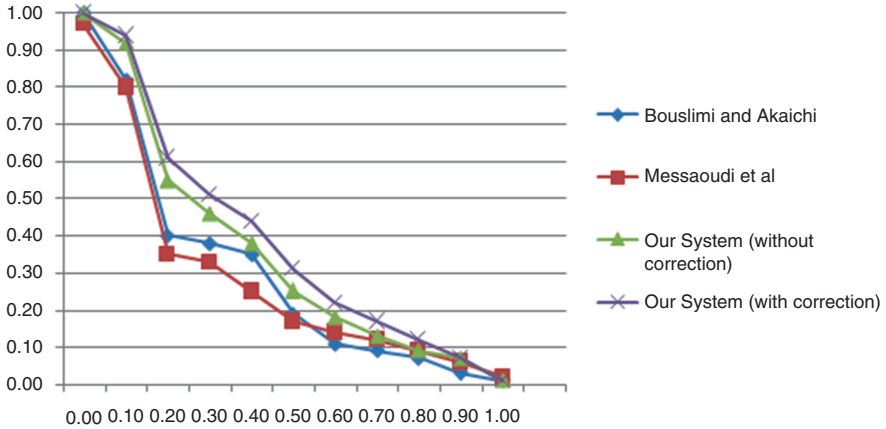


Fig. 2.9 Precision versus recall curves for corpus “transfCli-bio-txt”

6 Conclusion and Future Work

Social image indexing and retrieval in the large databases of the social networks advanced the challenges to form a new problem that needs special handling. Comments present a source for the indexation of images existing in the social network site. Firstly, we proposed in this paper a medical social network destined to both medical images presented by patients and physicians’ analyses expressing their medical recommendations and advice. Secondly, this study has presented a new approach to automatic multilingual annotation of medical images based on comments from the medical social network site. This approach has the primary

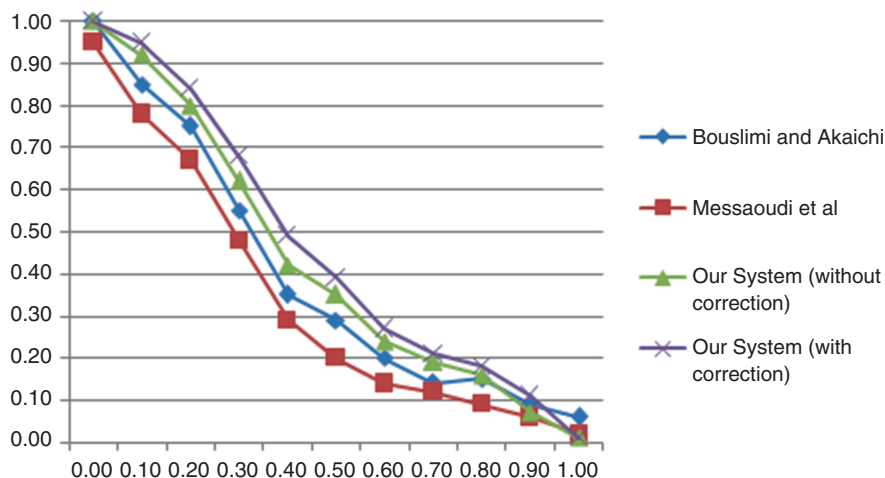


Fig. 2.10 Precision versus recall curves for corpus “AnnCardAng-txt”

goal of removing the ambiguity that comments are expressed in different languages. Our approach is about the extraction of multilingual terms and current concepts existing on comments, in order to facility the search task later. This approach focuses on algorithm mainly based on statistical methods and an external semantic resource. Statistical methods used to select important and significant multilingual terms from comments. Then an extraction of concepts step through a wealthy external multilingual medical semantic resource by a multilingual medical thesaurus (UMLS). Through the extraction of concepts, completeness is verified because it seeks to ensure that those selected keywords are the most complete possible. We worked also, in this paper, to remove the ambiguity of misspelled words in comments. The combination of the two metrics of Levenshtein and Jaccard is mainly the basic technique used in this study. We used a multilingual medical dictionary that contains the different vocabularies used in medicine. We evaluated with experiments on real annotations on medical images, which are obtained from the Hospital Charle-Nicolle in Tunis with collaboration between physicians from other countries. The results have shown the relevance of our approach, especially with European language. There are some limits in relation with other existing languages that need a specific treatment like Chinese and Arabic languages. Future work will focus on presenting a specific treatment for other existing languages in order to enlarge the circle of users of our social network site.

References

1. Zhi W, Wenwu Z, Peng C, Lifeng S, Shiqiang Y (2013) Social media recommendation. In: Social media retrieval. Computer communications and networks. Springer, Berlin. doi:10.1007/978-1-4471-4555-4 3
2. Doganay S (2014) Healthcare social networks: new choices for doctors, Patients. Available from <http://www.information.com/healthcare/patient-tools/healthcare-social-networks-new-choices-for-doctors-patients/d/d-id/1234884>
3. Franklin V, Greene S (2007) Sweet talk: a text messaging support system. *J Diabetes Nurs* 11(1):22–26
4. Grenier C (2003) The role of intermediate subject to understand the structuring of an organizational network of actors and technology case of a care network. In: Proceedings of the 9th conference of the association information and management, Grenoble
5. Messaoudi A, Bouslimi R, Akaichi J (2013) Indexing medical images based on collaborative experts reports. *Int J Comput Appl* (0975-887) 70(5):1–9
6. Daniel RG, Liza SR, Jennifer LK (2013) Dangers and opportunities for social media in medicine. *Clin Obstet Gynecol* 56(3). doi:10.1097/GRF.0b013e318297dc38
7. Feldman DL (2012) Medical social media networks: communicating across the virtual highway. *Q J Health Care Practice Risk Manag Infocus* 18(1):2–5
8. Maisonnasse L, Gaussier E, Chevallet J-P (2009) Combination of semantic analysis to search for medical information. In: RISE (Research Information semantics) within the INFORSID' conference, Toulouse
9. Gaussier E, Maisonnasse L, Chevallet JP (2008) Multiplying concept sources for graph modeling. In: CLEF 2007. LNCS 5152 proceedings, pp 585–592
10. Lacoste C, Chevallet JP, Lim j-h, Wei X, Raccoeanu D, Hoang D, Vuillenemot F (2006) Ipal knowledge-based medical image retrieval in imageCLEFmed 2006. In: Working notes for the CLEF 2006 workshop, Alicante
11. Neil S, Velte T, Jie H, Wei Z, Clement Y (2007) Knowledge intensive conceptual retrieval and passage extraction of biomedical literature. In: 30th annual 66 international ACM SIGIR conference on research and development in information retrieval
12. Li L-J, Fei-Fei L (2009) Optimol: automatic online picture collection via incremental model learning. *Int J Comput Vis* 88(2):147–168
13. Collins B, Deng J, Li K, Fei-Fei L (2008) Towards scalable dataset construction: an active learning approach. In: Proceedings of the European conference on computer vision
14. Mihalcea R, Leong C-W (2009) Towards communicating simple sentences using pictorial representations. *Mach Transl* 22:153–173
15. Von Ahn L, Dabbish L (2004) Labeling images with a computer game. In: Proceedings of the SIGCHI conference on human factors in computing systems, Vienna. ACM, New York, pp 319–326
16. Truran M, Goulding J, Ashman H (2005) Co-active intelligence for image retrieval. In Proc. of the 13th annual ACM international conference on multimedia, Hilton. ACM, New York, pp 547–550
17. Li Q, Lu SCY (2008) Collaborative tagging applications and approaches. *IEEE Multimed* 15(3):14–21
18. Shevade B, Sundaram H, Xie L (2007) Modeling personal and social network context for event annotation in images. In: Proceedings of the 7th ACM/IEEE-CS joint conference on digital libraries, Vancouver, BC. ACM, New York, pp 127–134
19. Stone Z, Zickler T, Darrell T (2008) Autotagging facebook: social network context improves photo annotation. In: Proceedings of the 1st IEEE workshop on internet vision (CVPR 2008), p 8
20. Bouslimi R, Messaoudi A, Akaichi J (2013) Using a bag of words for automatic medical image annotation with a latent semantic. *Int J Artif Intell Appl* 4(3):51

21. Barnard K, Forsyth D (2007) Learning the semantics of words and pictures. In: Proceedings of international conference on computer vision
22. Jeon J, Lavrenko V, Manmatha R (2007) Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the ACM SIGIR conference on research and development in information retrieval
23. Makadia A, Pavlovic V, Kumar S (2008) A new baseline for image annotation. In: Proceedings of the European conference on computer vision
24. Wang C, Blei David, Fei-Fei Li (2009) Simultaneous image classification and annotation. In: Proceedings of the IEEE conference on computer vision and pattern recognition
25. Zunjarwad A, Sundaram H, Xie L (2007) Contextual wisdom: social relations and correlations for multimedia event annotation. In: Proceedings of the 15th international conference on multimedia, Augsburg. ACM, New York, pp 615–624
26. Manning CD, Raghavan P, Schütze H (2008) Introduction to information retrieval. Cambridge University Press, New York
27. Fuming S, Yong G, Dongxia W, Xueming W (2010). A collaborative approach for image annotation. In: PSIVT, 2010, image and video technology, Pacific-Rim symposium on, image and video technology, Pacific-Rim symposium on 2010, pp 192–196. doi:10.1109/PSIVT.2010.39
28. Sun F, Ge Y, Wang D, Wang X (2010) A collaborative approach for image annotation. In: Proceedings of the PSIVT'10. IEEE Computer Society 2010, Singapore, pp 192–196. ISBN:978-0-7695-4285-0
29. Kanishcheva O, Angelova G (2015) A pipeline approach to image auto-tagging refinement. In: BCI '15 proceedings of the 7th Balkan conference on informatics conference, New York, NY. doi:10.1145/2801081.2801108
30. Bouslimi R, Akaichi J (2015) Automatic medical image annotation on social network of physician collaboration. *Netw Model Anal Health Inform Bioinforma* 4:10. doi:10.1007/s13721-015-0082-5
31. Harrathi F (2010) Extraction de concepts et de relations entre concepts à partir des documents multilingues: approche statistique et ontologique. PhD Thesis, INSA Lyon
32. Gong J, Sun S (2011) Individual doctor recommendation model on medical social network. In: Proceedings of the 7th international conference on advanced data mining and applications (ADMA'11)
33. Almansoori W, Zarour O, Jarada TN, Karampales P, Rokne J, Alhaji R (2011) Applications of social network construction and analysis in the medical referral process. In: Proceedings of the 2011 IEEE ninth international conference on dependable, autonomous and secure computing (DASC'11)
34. Xie Y, Chen Z, Cheng Y, Zhang K, Agrawal A, Liao WK, Choudhary A (2013) Detecting and tracking disease outbreaks by mining social media data. In: Proceedings of the twenty-third international joint conference on artificial intelligence (IJCAI'13)
35. Li J (2014) Data protection in healthcare social networks. *J IEEE Softw* 31(1):46–53
36. AMA Policy (2012) Professionalism in the use of social media. American Medical Association, 2012 Annual meeting. <http://www.ama-assn.org/ama/pub/meeting/professionalism-social-media.shtml>
37. Levenshtein VI (1966) Binary codes capable of correcting deletions, insertions and reversals. *Sov Phys Dokl* 10:707–710
38. Jaccard P (1901) Distribution de la flore alpine dans le Bassin des Drouces et dans quelques regions voisines. *Bull Soc Vaud Sci Nat* 37(140):241–272
39. Heasoo H, Lauw Hady W, Getoor L, Ntoulas A (2012) Organizing user search histories. *IEEE Trans J Mag Knowl Data Eng* 24:912–925
40. Navarro G (2001) A guided tour to approximate string matching. *ACM Comput Surv* 33(1):31–88
41. Frakes WB, Fox CJ (2003) Strength and similarity of affix removal stemming algorithms. In: Newsletter of ACM SIGIR forum homepage archive, vol 37(1), New York, pp 26–30
42. Jones K (1972) A statistical interpretation of term specificity and its application in retrieval. *J Doc* 28(1):11–21

43. Paukkeri M, Honkela T (2010) Likey: unsupervised language-independent keyphrase extraction. In: Proceedings of the 5th international workshop on semantic evaluation, Uppsala, Sweden, pp 162–165
44. NLM (2009) NLM unified medical language system fact sheet. Available from: <http://www.nlm.nih.gov/pubs/factsheets/umls.html>. Cited 23/04/2009

Prediction and Inference from Social Networks and
Social Media

Kawash, J.; Diaz, S.; Özyer, T. (Eds.)

2017, IX, 225 p. 82 illus., 54 illus. in color., Hardcover

ISBN: 978-3-319-51048-4