Exploring intentions on electronic health records retrieval

Studies with collaborative scenarios

Rodrigo Bonacin[1,3], **Julio Cesar dos Reis**[2], **Edemar Mendes Perciani**[3], **Olga Nabuco**[1]

1. CTI Renato Archer Campinas
São Paulo, Brazil
rodrigo.bonacin@cti.gov.br
olga.nabuco@cti.gov.br

2. Institute of Computing University of Campinas Campinas
São Paulo, Brazil
jreis@ic.unicamp.br

3. Faculty of Campo Limpo Paulista Campo Limpo Paulista
São Paulo, Brazil
edemar.mendes.perciani@gmail.com

ABSTRACT. *Despite the potential benefits of Electronic Health Records (EHRs), health care pro fessionals face difficulties in the selection of relevant documents in huge repositories during collaborative activities. In this article, we investigate the development of an innovative Information Retrieval (IR) and sharing mechanism that explores the formal representation of inten- tions in EHRs. To this end, this research relies on Organizational Semiotics and Speech Acts Theory. We defined an algorithm to filter and sort search results relying on intention classes explicitly declared as query parameters in the search mechanism. As our main contribution, we developed the SiRBI IR system for supporting group knowledge sharing through EHRs. To evaluate the proposal, we conducted an experimental study using a realworld EHR repository in two search scenarios, which involve an interdisciplinary group. The obtained results demonstrated the effectiveness of the solution.*

RÉSUMÉ. *Indépendamment des aspects positifs apportés par les dossiers médicaux partagés (DMP) informatisés, les professionnels de santé sont confrontés à des difficultés dans la sélection des documents pertinents, surtout dans la cadre des grandes bases de données lors des activités*

Ingénierie des systèmes d'information – n° 2/2018, 111-135

de collaboration. Dans le cadre de cet article, nous nous sommes appuyés sur le développement d'un mécanisme innovant de Recherche d'Information (RI) qui explore la représentation formelle des intentions dans les DMP. Cette recherche repose sur la théorie organisationnelle de la sémiotique et de la théorie des actes de langage afin de catégoriser plusieurs types d'intentions. Notre étude porte sur des problèmes de définition, sélection et classement des résultats de recherche et nous examinons les intentions explicitement déclarées par les utilisateurs. Notre principale contribution est le développement d'un système RI qui vient en appui au partage des connaissances de groupe via DMP. Pour évaluer cette proposition, nous avons mené une étude expérimentale d'après un référentiel DMP réel de dossiers médicaux. Deux scénarios sont définis qui impliquent un groupe interdisciplinaire de professionnels de santé. Les résultats obtenus sont analysés à partir de mesures comme la précision et le rappel et ont démontré l'efficacité de cette solution.

KEYWORDS: *information retrieval, electronic health records, information sharing, query expansion, intentions, illocutions, speech acts theory.*

MOTS-CLÉS : *récupération de l'information, dossier médical électronique, expansion de requêtes, les intentions, la théorie des actes de langage.*

DOI:10.3166/ISI.23.2.111-135

1. Introduction

Nowadays, we are experiencing the growth of a new area of communication, and it is up to us to explore the most positive potentials of this area (Lévy, 2001). As argued in Khan *et al*. (2010), health "information fusion" and social networks can lead to a revolution in public health. New health treatments, for instance, can be informed by technologies that facilitate collective intelligence within health practice communities (Hesse *et al*., 2010). In this context, the creation of communities of researchers and health professionals is a process that frequently faces complex issues. Collective intelligence helps to better take into account the intentional aspects of communication among members.

Electronic Health Records (EHRs) are valuable tools for clinical research communication and information sharing, improving the research of new treatments and care dedicated to patients (Chen *et al*., 2012). Hesse *et al*. (2011), for instance, emphasize that the adoption of interoperable and data-reliant EHRs remains a change driven by the health care sector towards "community intelligence". Information Retrieval (IR) tools can be explored to improve collaboration and information sharing within communities of health care professionals and researchers. However, a significant amount of clinical documentation in EHRs is unstructured, in narrative format. This format allows clinicians to express their thoughts and to write rich patient's stories (Zheng *et al*., 2011).

In general, IR techniques have explored syntactic-based approaches, as well as methods tackling the meaning of terms (Gurulingappa *et al*., 2011; Dong *et al*., 2008). Nevertheless, challenges remain due to the difficulties of dealing with non-structured

texts in Natural Language (NL) with specific aspects of medical information. More specifically, there is lack of studies focused on addressing issues related to: (1) how to support the users so that they make their intentions explicit in IR processes; and (2) how to define algorithms that consider the users' declared intentions during the selection and ranking of search results from EHR repositories.

The use of intentions refers to a central aspect of human communication because it helps in the interpretation and understanding of shared information. The ability to recognize and understand the others' intentions is central in "Theory of Mind" (ToM) (Apperly, 2010). This ability affects the social sensibility of a group's members, and consequently, collective intelligence.

We argue that IR techniques might benefit from explicitly using declared intentions to disambiguate terms within a delimited context. For instance, the relevance of medical documents in IR might change if physicians intended to describe or prescribe a specific treatment in a given health context. Thus, this is a tool for supporting information sharing and recovering in collaborative interdisciplinary groups, by taking into account one key aspect of the social sensibility; *i.e.*, this tool incorporates the ability to recognize and understand the others' intentions so as to support group knowledge sharing.

The following scenario illustrates the difficulties involved in this research. A physician may want to recover all EHRs in which patients present a cough for more than four weeks, and deny any allergy. In this scenario, a syntactic-based IR mechanism might not be able to recover related terms to cough and/or allergy. This issue is potentially solved by semantic IR methods, but they still may be unable to relate the temporal aspect (more than four weeks) with the specific concepts (presence of a cough and the denial of an allergy). These methods are unable to consider the intention of whom wrote the text to express and share with others the presence of a cough and the absence of an allergy. Furthermore, there is a lack of models to categorize these dimensions, *i.e.*, distinguish the intention from a domain term. In many cases, the terms that identify intentions are treated as stop-words and are removed from the query terms.

This research addresses the problem of improving communication by proposing IR methods in text-based fields of EHRs. Regarding this issue, we assume the existence of an EHR repository consisting of a finite set of medical documents with texts to be retrieved by an interdisciplinary team. Each document contains NL texts with medical data, *e.g.*, within a case's history (Hx or anamnesis).

In this article, we investigate alternatives for combining keywords with types and characteristics of intentions. In the search process, we match the declared intentions in a query with the intentions annotated in the documents. In addition, this problem demands suitable solutions with positive effects when a type of intention is included in the search query.

We define the search method by relying on the annotation of meanings and categories of intentions in NL descriptions present in EHRs (*e.g.*, in medical histories).

The technique explores the definition of terms in medical texts based on Knowledge Organization Systems (KOS). The proposal combines the annotated meanings with the type of intention expressed in fragments of text in EHR documents.

This research proposes and formalizes the *PraSA* – (Pragmatic Search Algorithm) as a ranking algorithm to properly sort the search results according to user's search input strings, which express his/her intentions. Our proposal depicts the selection and best ordering of a subset of documents from the original EHR repository according to the given query and its parameters. We demonstrate the way in which the declared intention specified in the user's query matches the annotated intention in the EHR documents. We implemented the *SiRBI* – (System for Information Retrieval Based on Intentions) as proof of concept in this research.

The solution involves dimensions representing intention classes in real patients' EHRs data. The theoretical basis of the proposal includes Speech Act Theory (Searle, 1969) and Semiotics (Peirce, 1931–1935; Liu, Li, 2014). These theories and studies provide useful methods to identify and to structure aspects related to intentions in communication acts.

In this research, we conducted an experimental validation by exploring two distinct scenarios defined by health care professionals. The hypothesis is that by using intentions the effectiveness of the search results will improve. The obtained findings reveal the applicability of the method in search scenarios in EHRs repositories. These scenarios involve collaboration among multiple health professionals, including scenarios defined by a physician and 4 nurses. These scenarios were elaborated to explore intensive information sharing among members of multidisciplinary teams of health professionals. The effectiveness of the approach was determined using a set of medical documents. Our tool was not limited by lexical-syntactic aspects of the content in the documents, and retrieved different documents with high scores of precision and recall.

The remaining of this article is organized as follows: Section 2 presents background information, such as fundamental concepts, our adopted theoretical framework and related work. Section 3 describes our Illocution-driven information retrieval method with the proposed algorithm and system. Section 4 details the conducted experimental evaluation; Section 5 discusses the obtained findings. Finally, Section 6 wraps up the article and points to future investigations.

2. Background

This section presents the fundamental concepts used in this paper (2.1), followed by the adopted theoretical framework (2.2), and related work (2.3).

2.1. Fundamental Concepts

2.1.1. EHRs, Semantic IR and Query Expansion

Medical records, since the ancient Greek physician Hippocrates, contain medical information grouped and stored chronologically, showing the evolution of a disease in time (Totelin, 2009). In the last decades, medical records have been digitized and inserted into computational systems in order to target better patient outcomes and prevent adverse effect of care. According to Laforest and Tchounikine (1999), EHRs are the main tool used to centralize and coordinate medical research. Several health professionals access and fill-out this tool. Nowadays, leading organizations such as *OpenEHR* and Health Level Seven (HL7) proposed international standards for EHRs.

OpenEHR is an open domain-driven platform for developing flexible e-health systems. Its architecture approach distinguishes data from a model, enabling the building of a repository of EHRs. The models, known as archetypes, are sharable and reusable in different projects. *OpenEHR* plays a key role in standardization initiatives, which can improve data traceability and, therefore, support medical IR. However, its focus is mainly on quantitative aspects, retrieving only pre-stablished archetypes. The qualitative information is usually consisted of free-text written in NL describing the patient's status and diagnostics (Zheng *et al*., 2011).

Metadata can be used for providing additional structured information associated with EHR and other clinical documents. It can be applied, for instance, to improve the management of personal health records (*e.g.*, (Alyami *et al*., 2017)), in big data applications (*e.g.*, (Sweet, Moulaison, 2013)), as well as in IR mechanisms.

In this paper, we assume that the explored EHRs are well-structured according to the existing techniques and models. Our IR technique focuses on the free-text written in NL (*e.g.*, pre-consultation descriptions and anamnesis). More specifically, we aim to improve the relevance of search results based on "internal" textual elements of these models, and, therefore, improve information sharing and collaboration amongst health professionals. In our proposal, we explore metadata in IR mechanisms, including descriptions of semantics and intentions associated with free text in EHR.

Nowadays, a large volume of text-based documents must be accessible using IR mechanisms. Search engines are usually composed of three components: (1) a database containing indexes for documents; (2) query and results interfaces; and (3) algorithms capable of selecting the results. The difficulties to retrieve the correct content in non-structured content make users perform several attempts using different search (input) parameters. These difficulties are aggravated in IR mechanisms that search only lexical-syntactic comparisons. These techniques are usually not able to return relevant results due to the complexity of the terms' semantic, such as polysemy and synonymous words. Thus, this problem demands complex IR mechanisms, including the use of artificial intelligence and semantic web techniques (Aroma, Kurian, 2013).

A semantic IR process can be summarized in three main phases, as exemplified: (1) natural language data interpretation, where relevant concepts in the sentence are

extracted; (2) a set of concepts used in queries (input) are confronted with artifacts that represent domain knowledge; and (3) finally, the results are presented to the users in the computational system's interface (Hildebrand *et al*., 2007).

Semantic IR mostly makes use of Semantic Web technologies (Guha *et al*., 2003), such as ontologies, to retrieve information more effectively when considering the meaning of available data. The expansion of queries is a semantic IR technique in which it is possible to detect and exploit related terms to the originally inserted keywords typed by the user.

Query Expansion can play a central role in the tool's efficiency because the user's vocabulary for a query topic is generally less diversified than the domain vocabulary (Chawla, Bedi, 2008). In addition, query results often do not reflect users' needs due to the impact of various language phenomena that hinder the search. In multidisciplinary teams, the participants' different experiences and backgrounds aggravate this.

According to Sharef and Madzin (2012), query expansion can extend the original query using extra knowledge representation. This allows for the search results to be more meaningful with respect to domain documents. Essentially, the process benefits from knowledge bases to select new terms that compose the query in its new version. The knowledge bases define alternative concept labels to refine meanings, as well as relationships between different concepts (Chawla, Bedi, 2008).

2.1.2. Unified Medical Language System (UMLS)

In this investigation, we adopt the UMLS as a knowledge source to explore the query expansion technique. The UMLS (Bodenreider, 2004) is a project developed by the NLM (*National Library of Medicine of the United States*) since 1986. It consists of an extensive terminological library that combines approximately 200 domain-specific vocabularies, therefore representing biomedical knowledge in several languages.

UMLS consists of three major components with the purpose of forming a unified structure for different sources of knowledge in the biomedical area: (1) the **semantic network** classifies concepts into semantic types, and establishes relationships between them; (2) the **metathesaurus** aggregates concepts from several sources and describes the definition of the concepts; (3) the **knowledge bases** refer to a set of controlled vocabularies and international standards with lexicographic information (*e.g.*, MeSH, SNOMED CT, ICD, NCI, and LOINC).

We use the UMLS as part of our solution because it provides a model for knowledge representation that allows for the exploration of semantic relationships present in the medical domain and, thus, allows for the use of query extension techniques.

A keyword expressed in a user query can be searched in UMLS to retrieve its meaning, related concepts, and associated semantic types. For example, a user who looks for "*Heart*" in an EHR repository may also be interested in EHRs that contain related terms in their descriptions, such as EHRs reporting problems in the "*Endocardium*" and "*Heart atriums*". Furthermore, the UMLS contains vocabularies in the

Portuguese language, which enables its use in the context of the EHRs explored in our experimental evaluations (*cf.* Section 4).

2.2. Explored Theoretical Framework

Semiotic is the science that studies signs (Peirce, 1931–1935). In semiotics, according to Morris (1937), syntax studies the interrelation of signs; semantics studies relations of the signs and the objects; and pragmatics studies the relations of signs to their interpreters. Pragmatics defines intentional behavior (Liu, 2000; Liu, Li, 2014), where the context is used to indicate intention. The central element of this work considers pragmatics as means to improve IR and sharing, going beyond technologies that consider syntactic and semantics aspects.

In addition to the concept of pragmatics, as defined in the semiotics field, this investigation has its basis on the Speech Act Theory (SAT). This theory emerged in the mid-fifties, having as its pioneers Austin (1962), followed by Searle (1969). They questioned the passive role of the language, used to describe the state of things. Austin has shown that certain statements are not meant to describe anything, but rather to perform actions. According to Austin, speech acts are constituted by three dimensions, respectively: locutionary (performing an utterance), illocutionary, and perlocutionary. An illocutionary act is the attempt to communicate in the sense of expressing an attitude, an illocution is what was meant in a speech act, and it carries the speaker's intentions. A perlocutionary act is the actual effect produced in the addressee. In this work, we focus on locutionary and illocutionary acts. Searle (1969) proposed a classification of speech acts, which includes the following classes: assertives, directives, commissives, expressives, and declarations. In addition to Austin's and Searle's frameworks, several alternative frameworks appeared to categorize the speech acts.

Communication can be understood as an intentional system in which humans act and interact with one another to achieve goals that may be related to the community or its individuals (Liu, 2000; Liu, Li, 2014). Based on Semiotic and Speech Acts, Liu (2000) defined a communication act as a ternary structure consisting of the speaker, the receiver, and the message. The message can be divided into two parts: the content and the function. The content aggregates the meaning of the message expressed in the proposition, and the function specifies the illocutions through signs, reflecting the speaker's intention in the message. This theoretical framework allows us to link the content with the function by using, for example, ontologies (Bonacin *et al.*, 2012). Therefore, this framework provides means for the development of IR mechanisms.

Liu (2000) proposed a framework where Illocutions are grouped into three dimensions. In one dimension a distinction is made between descriptive and prescriptive "inventions"; another distinction between affective and denotative "modes"; and finally to distinguish different "times", such as future or present/past. If an illocution, in a communication act, tends to express the emotional state of the announcer, this is classified as affective. Otherwise it is classified as denotative. If an illocution has the function of expressing an inventive or instructive effect, it is a prescriptive illocu-

tion. Otherwise it is a descriptive one. By analyzing these three dimensions, we have an illocution classification framework (Liu, 2000; Liu, Li, 2014), with eight classes, described as follows:

1. **Proposal**: This class refers to illocutions that are in the future (time), prescriptive (invention) and denotative (mode), e.g., request something, give a command or promise/guarantee;

2. **Inducement**: This class refers to illocutions that are in the future (time), prescriptive (invention) and affective (mode), e.g., threaten, warn or tempt someone;

3. **Forecast**: This class refers to illocutions that are in the future (time), descriptive (invention) and denotative (mode), e.g., predict or assume something;

4. **Wish**: This class refers to illocutions that are in the future (time), descriptive (invention) and affective (mode), e.g., wish, hope and desire something;

5. **Palinode**: This class refers to illocutions that are in the present/past (time), prescriptive (invention) and denotative (mode), e.g., retract something, annul or revoke an act;

6. **Contrition**: This class refers to illocutions that are in the present/past (time), prescriptive (invention) and affective (mode), e.g., regret or apologize;

7. **Assertion**: This class refers to illocutions that are in the present/past (time), descriptive (invention) and denotative (mode), e.g., make an assertion, report or notify;

8. **Valuation**: This class refers to illocutions that are in the present/past (time), descriptive (invention) and affective (mode), e.g., judge, complain or accuse.

2.3. Information Retrieval based on Intentions

We performed a literature review of studies that address intention in IR, as well as IR techniques applied to EHR repositories. We briefly describe our review procedure as follows. First, we searched scientific bases (ACM Digital Library, IEEE Xplore, Springer Link, ScienceDirect, and Google Scholar) using the strings presented in Table 1. The review was performed during May of 2016 without filters for dates. The next step was to read the papers' titles, abstracts, and keywords in order to select related studies for full reading and analysis in pairs. Finally, we selected, according to these studies' contributions, a small group of papers to be discussed in this article. In this last step, 18 papers were selected to be presented from a total of 43 papers.

The existing studies explored intentions in IR techniques in various contexts and domains. Hwang *et al*. (2011), for instance, proposed a model for IR in which user's intentions are detected through sensors in a pervasive computing environment. By using sensors, a computational model identifies the user's intentions at the search time. Gupta *et al*. (2012) presented a search model using intelligent agents and filters that aim to capture the user's intentions by using results from Google's search engine. In their approach, the users assign checkboxes associated with the results from Google, which are used in the next searches.

Table 1. Search Strings for the literature review

Intention aspects in IR	(“Information retrieval” AND Intention) OR (“Information recovery” AND Intention) OR (“Information retrieval” AND Intended) OR (“Information retrieval” AND Purpose) OR (“Information recovery” AND Purpose) OR (“Semantic search” AND Intended) OR (“Semantic search” AND Intention) OR (“Semantic search” AND Purpose)
IR techniques applied to EHR	(“Information retrieval” AND “Electronic medical records”) OR (“Information retrieval” AND EHR) OR (“Information recovery” AND “Electronic medical records”) OR (“Information recovery” AND EHR) OR (“Information retrieval” AND “Medical information”) OR (“Information recovery” AND “Medical information”) OR (“Semantic search” AND EHR) OR (“Semantic search” AND “Information medical”)

Some investigations have considered the user's effort as a central issue in intention based IR; Mendoza and Baeza-Yates (2008), for instance, proposed an automatic method for categorizing the behavior of queries performed by users. The algorithm aims to understand the user's intentions by analyzing the keywords typed by users. Tang *et al.* (2012) proposed an approach that searches images by using keywords expansion, and visual similarity.

In another proposal, Noor and Martinez (2009) explored social data for improving the IR process. The study defined a user interest model that serves as an interpretation of the user's intention. Zinglé (2006) presented a search engine based on conceptual graphs that represents concepts related to a domain. The concatenation of lexical-syntactic expressions, semantic knowledge, and representation of pragmatic aspects allow users to make their intentions explicit in the query.

Few studies have emphasized the intentions of those who produced the contents. Asai and Yamana (2014), for instance, presented a framework for improving the learning of annotation systems. It detects intentions through annotation in paper-based documents, and it is able to highlight a range of content in a document, extract search terms, and provide comments on specific content. Gómez *et al.* (1999) proposed a mechanism of IR that aims to determine the intention by analyzing the titles and summaries in a domain-independent approach. This mechanism is based on pattern match, and the identification of pre-defined verbs and constructs.

We also emphasize techniques that explore semantic and pragmatics aspects of IR in EHR repositories. Tawfik *et al.* (2011) presented an experiment that compares patient-specific information and semantic research results from a human-computer interaction viewpoint. Their results suggested that the semantic approach can reduce cognitive load within clinical settings according to the similarity of patient-specific information needs. Besides, Koopman *et al.* (2016) showed that semantic inference can retrieve many new relevant documents that are not retrieved by state-of-the-art IR.

Popescu (2010) explored fuzzy rules in ontological bases to recover similar treatments or synonymous terms. For a given diagnosis, the user can consult treatments that were well succeed or not. By using machine learning techniques, Cogley *et al.* (2013) presented an approach that is able to extract domain entities from unstructured

documents. NL queries are then transformed into a structured format. Kreuzthaler *et al*. (2015) proposed a browser-based platform for document import, search, and navigation using NL processing and semantic-enriched medical texts.

Standards such as UMLS and HL7 were explored by various studies on IR of medical data. Gurulingappa *et al*. (2011) proposed a platform for IR of EHR that explores manual keyword extraction, semantic search using UMLS, and domain Ontologies. Perez-Rey *et al*. (2012) developed a mechanism to facilitate the construction of queries to retrieve scientific literature related to EHRs. Bhatt *et al*. (2009) presented a technique for creating sub-ontologies (e.g., by medical specialty) using semantic web standards (RDFs and OWL), domain standards, and vocabularies encompassed by the UMLS. Pruski and Wisniewski (2012) presented an approach for IR in EHRs containing shared encrypted clinical documents that parses UMLS properties before encoding the content.

Decision support and communication are other issues to be considered in order to provide better IR. Liaw *et al*. (2014), for instance, presented an ontology-based approach for integrating EHRs to improve decision support and quality of care. Forcher *et al*. (2014) proposed the development of a semantic mechanism that provides intuitive justifications of medical semantic search results based on Wikipedia articles.

Several studies reported positive experiences in IR techniques applied to EHR and the use of intentions IR. However, there is still a lack of techniques that consider the relations between meanings and intentions in EHR free-text content. This article presents an original method, an algorithm, and a system prototype for IR based on illocution. The intentions are represented in the produced content via explicit annotations; and in queries via structured parameters. This study aims to improve information sharing and collaboration in multidisciplinary groups, by providing a tool that supports the searcher' ability to recognize the intentions of those who produced the EHRs.

3. Method and Algorithm Conceptualization

In this section, we first present the description of our method for Illocution-driven IR (subsection 3.1). We then present the *PraSA* algorithm (subsection 3.2), and the *SiRBI* system (subsection 3.3), which was implemented to explore the potentialities of the proposed method and algorithm.

3.1. Illocution-driven IR Method

3.1.1. Analyzing Intentions in EHR

This work started with the analysis of existing EHRs to investigate the way that health care professionals use domain terms for expressing intentions. This step was useful for informing decisions in the subsequent elements of this investigation. This research considered 10.200 EHRs available in a public hospital from the city of "Aguas de Lindoia" in Brazil. We manually analyzed a subset with 26 "dengue fever" cases

with free-text notations of pre-consultation and the patient's history. This analysis was performed in 5 steps (*cf.* Figure 1) with the participation of a physician.

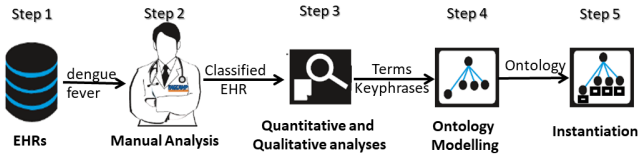

Figure 1. Intention Analysis

In the first step, the EHRs were selected according to a diagnosis hypothesis (*i.e.*, dengue fever). In the second step, the researchers, with the support of a physician, classified the EHRs using Liu's framework. A qualitative and quantitative analysis of the illocutions were performed in the third step; which was then used in the forth step for the modeling of an ontology to represent the illocution classes and concepts. Finally, we instantiated the model with the analyzed content in the fifth step. Details of the procedure and results are presented in Reis *et al.* (2016b). This study revealed that the analyzed texts are concise and impersonal. This was observed, for instance, by the high incidence of assertions. The most relevant aspect of this analysis identified the terms frequently used to express each illocution dimension and class.

These observations were relevant for the definition of the query strategies employed in our IR method. We assumed that search parameters related to the illocution dimensions may contribute to improve the precision of the search results. Using the dimension time, for instance, it is possible to distinguish coughing for one day as a different indicator from coughing for one month. We explicitly used these aspects in the query formulation phase in our method.

3.1.2. Description of the method

Figure 2 presents an overview of the proposed solution. EHRs are stored in a repository (item D of Figure 2) annotated with semantically described concepts (item E of Figure 2) and the type of illocutions (item C of Figure 2).

The *PraSA* algorithm (item G of Figure 2) (*cf.* subsection 3.2) has as input of the following elements: (i) the EHRs from the repository to be searched (item D of Figure 2); (ii) the metadata as the semantically annotated concepts of the EHRs according to a KOS (item F of Figure 2); (iii) the domain concepts from the KOS; and, (iv) a set of keywords and the type of illocutions specified by the user (item A of Figure 2). As results of its execution, a sorted list of EHRs (item H of Figure 2) is provided. In the following paragraphs, we detail each element of the solution.

Semantic annotation (item E of Figure 2) is based on the inspection of KOS concepts which are linked with text fragments of the EHRs (item D of Figure 2). By relating text fragments with KOS concept codes, the method generates semantic metadata for the medical records. For example, in the following text fragment "*The patient is under treatment for asthma for seven days*", the term "*asthma*" is detected as a concept: "*Asthma (disease)*" with code: "*SCTID: 195967001*" in the SNOMED CT.

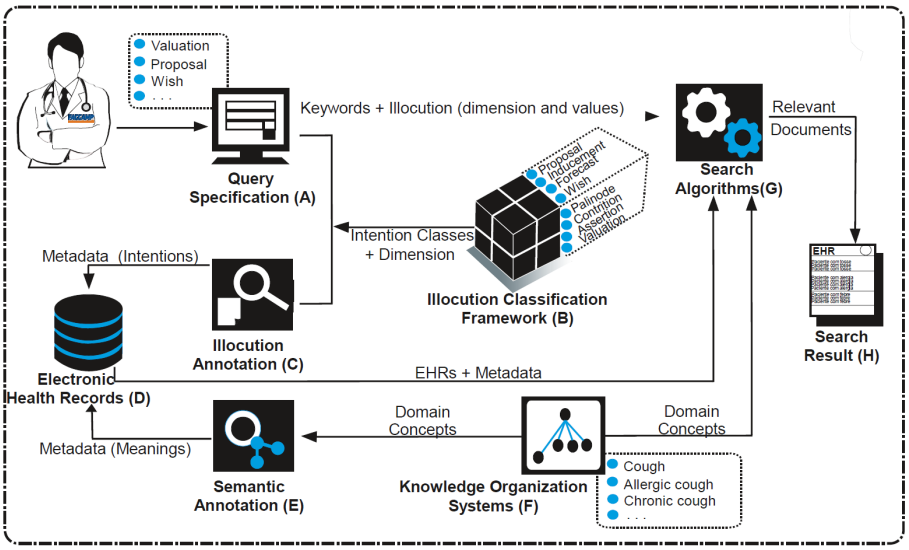

Figure 2. Overview of the Information Recovery Method

In this study, we explored existing solutions for the semantic annotation task. However, there was the need to deal with the constraints of having a Portuguese language EHR repository. In our solution, we defined a procedure that makes requests to a set of services and programming patterns for accessing the UMLS. We explored the Portuguese version of the KOSs integrated into the UMLS (item F of Figure 2).

Items B and C in Figure 2 show an overview of the adopted process to annotate the illocutions, *i.e.*, the process of linking text fragments of the EHR content to illocution dimensions. For instance, the following texts "... *patient refers to fever, headache, retro eye pain and myalgia ....*" and "... *deny allergy and commodity ...*" can be annotated as "assertions" (time: present/past, invention: descriptive, and mode: denotative), whereas the following text "... *request complete HMG, paracetamol ...*" can be classified as a "proposal" (time: future, invention: prescriptive and mode: denotative).

The combination of results of semantic annotation with annotation of illocutions refers to an original key aspect of this research. Based on the study presented in subsection 3.1.1, we defined and implemented an algorithm to automatically detect illocutions in free-text in the Portuguese language (Reis *et al.*, 2017). This algorithm obtained *0.92* of *F-measure* using a "dengue fever" EHR dataset. In addition, we developed a tool for manual annotation of illocutions in EHRs (*cf.* subsection 3.3).

Figure 3 details our approach for the search query definition. The search string includes: (i) a set of keywords related to diseases, symptoms, or any other relevant terms; (ii) an illocution dimension associated with a keyword; and (iii) values associated to such illocution. For instance, **Keyword**: "Cough" $\wedge$ **Dimension**: $< Time >$

$\wedge$ **Value**: "more than 4 weeks", means that the user is searching for EHRs concerning patients that reported a cough for more than 4 weeks.

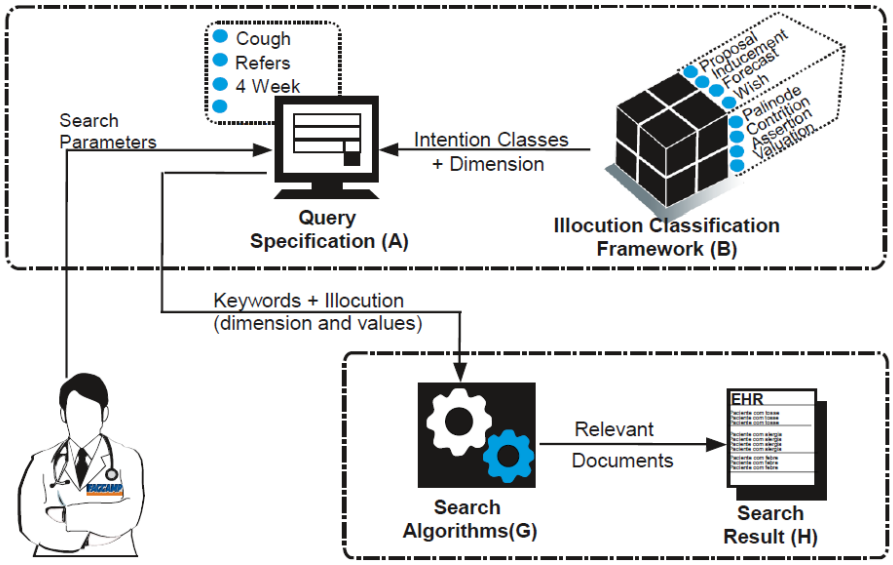

Figure 3. Search string specification

3.2. PraSA – Pragmatic Search Algorithm

In our method, the *PraSA* (Algorithm 1) is used to filter and rank the results according to the informed search string. It receives the following input: R defines a set of documents $d_{i,i\in N}, R = \{d_1, d_2, ..., d_n\}$. The set S is defined from the KOS concepts $c_{i,i\in N}, S = \{c_1, c_2, ..., c_k\}$. The search string is defined as a set of elements $C = \{E_1, E_2, ..., E_x\}$ where $E_i = (p_i, dm_i, vl_i)$ is a triple with a keyword (p), a dimension (dm), and an illocution value (vl). The metadata is defined by a set $M = \{A_1, A_2, ..., A_y\}$, where $A_{y,y\in N}$ is a $4-tuple$ with the elements $(d_i, st_{i,j}, dm_i, vl_i)$, to which $st_{i,j,i\in N,j\in N}$ refers to the jth sentence st_j in the document d_i, $st_{i,j} \in d_i$.

The *PraSA* begins with the syntactic retrieval of all documents where there exists an occurrence of each keyword informed by the user (lines 4-6 of Algorithm 1). The results are stored in REL_{INI}. Afterwards, the algorithm checks for the occurrence of the keywords in the entire KOS content (line 7 of Algorithm 1). Then, it performs the expanded search for each related term (lines 10-12 of Algorithm 1). These results are stored in REL_{EXP}. In the sequence, it makes the union of REL_{INI} and REL_{EXP} (line 13 of Algorithm 1).

In next step, it performs the selection of results based on the annotations (lines 14-16 of Algorithm 1). At this stage, the algorithm selects all EHRs from REL_U where

keywords, concept labels, or selected terms (synonyms and related terms) appear inside of text fragments annotated with the same input type of illocution and dimension values. Finally, the *PraSA* reorders the results according to the type of match between terms and illocutions occurrences found. The algorithm gives priority to documents that provide exact input keywords as to those of the query found in the illocution annotation. The algorithm's second priority are EHRs with synonyms and other terms related to annotated illocution (line 17 of Algorithm 1). We present in (Reis *et al.*, 2016a) an entire description of a usage scenario, and examples that illustrate the use of the *PraSA*. This scenario, which aims to recover all the patients' registers that reported on a cough for more than four weeks and deny any allergy, is summarized as follows: the *PraSA* (1) retrieves the registers with the keywords "cough" and "allergy"; (2) lists synonyms and related terms; (3) performs a semantic retrieval of these terms (*e.g.*, hypersensitivity and atop); (4) unifies the results; (5) filters the registers according to the annotations; and (6) reorders the registers according to the keywords and illocutions.

Require: R, S, M, C
Ensure: REL'_F *// Ordered set with the final result*
1: **Begin**
2: $REL_{INI} \leftarrow \emptyset$
3: $T \leftarrow \emptyset$
4: **for each** $E \in C$ **do**
5: **for each** $p \in E$ **do**
6: $REL_{INI} \leftarrow REL_{INI} \bigcup Synt\_SEARCH(p, R)$
7: $T \leftarrow T \bigcup EXPAND(p, S)$
8: **end for**
9: **end for**
10: **for each** $t \in T$ **do**
11: $REL_{EXP} \leftarrow REL_{EXP} \bigcup Synt\_SEARCH_{EXP}(t, R, T)$
12: **end for**
13: $REL_U \leftarrow REL_{INI} \bigcup REL_{EXP}$
14: **for each** $E \in C$ **do**
15: $REL_F \leftarrow SELECT(REL_U, E, T, M)$
16: **end for**
17: $REL'_F \leftarrow REORDER(REL_F, C, M)$
18: **return** REL'_F
19: **End**

Algorithm 1: Pragmatic Search Algorithm

3.3. SiRBI – System for Information Retrieval Based on Intentions

The *SiRBI* system implements our method reusing existing search tools (Apache

Solr[1]) and KOSs (UMLS). Figure 4 presents an overview of its architecture, which is composed of 7 components described as follows:

– **Term Analysis (item A)**. This component (pre)processes all the EHRs in the database to deal with special character encoding, eliminate stop-words and split free-text in keyphrases to be used in the expanded search;

– **Database Component (item B)**. Database with the (pre)processed EHRs;

– **Query Expansion Component (item C)**. This component accesses external KOSs for each keyphrase and returns a set of identification codes that are stored in the database component (item B of Figure 4); *i.e.*, it associates standardized terms returned from A with synonyms and related terms. This occurs in a preprocessing stage that aims to enable query expansion in execution time;

– **External KOSs (item D)**. In the current version, we access UMLS-UTS services[2] to retrieve external KOSs. Our approach is not limited to this service and additional KOSs can be aggregated;

– **Syntactic Search Component (item E)**. This component accesses an external tool in order to produce the indexing and syntactic search of all EHRs in the database. At this version of the implementation, we explore the Apache *Solr*;

– **External Syntactic Search (item F)**. External tool that enables the creation of an index regarding the EHR contents, and performs the syntactic search. In the current version, we implemented this component using the *Apache Solr* tool. *SiRBI* system is not restricted to this tool, and other IR tools can be used in our implementation;

– **Pragmatic Search Component (item G)**. This component provides the user interfaces for the specification of queries, and implements the *PraSA* algorithm. It receives the query parameters as input from the user. It explores the other components to perform the search, including: (1) the database access (item B); (2) the preprocessing data (by items A and C); (3) the synonyms and related terms as input parameters (access to item B at the query execution time); and (4) multiple queries submitted to item E (as required by *PraSA*).

In order to illustrate the query specification, Figure 5 presents the user interface developed in the *SiRBI*. Where users can specify the search parameters and perform the search by including keywords, dimensions, and dimension values. As shown in Figure 5, the system suggests standardized terms that can be used as keywords and dimension values. Users should understand the concepts of illocution dimensions and values to use the interface, this problem was minimized by using hints when the they point the mouse over a dimension of value.

1. lucene.apache.org/solr
2. https://uts.nlm.nih.gov/home.html

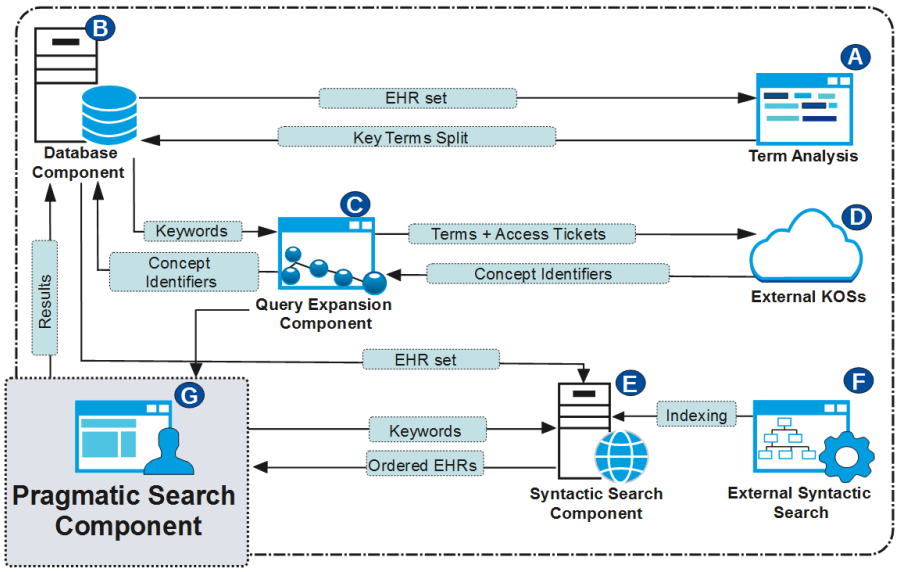

Figure 4. The SiRBI system Architecture

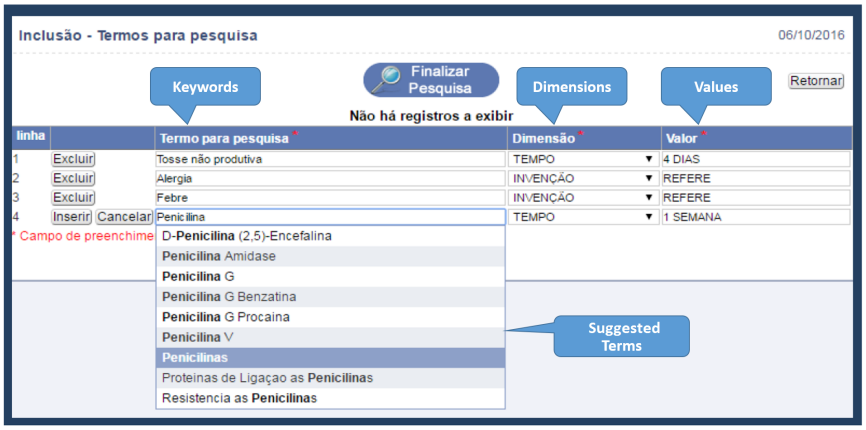

Figure 5. User interface for search query specification in the SiRBI system

4. Experimental Evaluation

Since there are no databases for measuring users' intentions ready to use, the experimental evaluation requires a lot of manual work. In this section, we first present the participants involved and experimental scenarios employed, with the help of a physician and 4 nurses, to simulate IR scenarios in interdisciplinary teams (subsection 4.1). We then describe the constructed dataset, which was used as a reference set to

test the search scenarios (subsection 4.2). Based on this scenario, we defined several configurations that were then tested to analyze the effectiveness of the *SiRBI* based on objective evaluation metrics (subsection 4.3). Subsection 4.4 details the obtained results.

4.1. Participants and Scenarios

This study was initially carried out in conjunction with two health professionals with different specialties and training. The first professional is a physician, an allergy specialist with 22 years of experience in the field. The second is a nurse specialized in gynecology and obstetrics, with 18 years of work-experience. These professionals, directly involved in the evaluation tasks, were assisted by 3 other nurses, which reviewed the evaluation of the first two professionals. The health professionals defined the relevant search scenarios and acted in the classification of EHRs to create our dataset of reference. The search results obtained with the developed algorithm were compared to the dataset of reference to measure effectiveness according to evaluation measures.

In order to define the search scenarios in conjunction with the physicians, an initial exploratory analysis was carried out using a total of 13.300 EHRs from a public hospital in the city of "*Aguas de Lindoia*", located in the State of São Paulo, Brazil. We explored the ICD codes associated to the EHRs containing a diagnostic hypothesis in order to select and restrict a subset of adequate EHRs to study. The analysis indicated the diagnostic hypotheses with higher incidence rates in the whole dataset. Based on diagnostic hypothesis, the health care professionals could elaborate questions about topics to which they would like to obtain more precise answers. Two scenarios are described as follows:

Scenario 1. In this scenario, the professionals would like to retrieve EHRs from patients who report a cough at night, but do not have any allergies. The detection of the persistence of a cough at night can help identify the causative agent (dust, fungi, mites, *etc*.) that settles and thrives in the residential environment. In this case, the medical conduct must include initiatives to combat the causative agents. This scenario was composed by a query that involved two sentences as follows:

1. **Keyword**: "Cough" $\wedge$ **Dimension**: $< Time >$ $\wedge$ **Value**: "At Night";
2. **Keyword**: "Allergy" $\wedge$ **Dimension**: $< Mode >$ $\wedge$ **Value**: "Deny";

Scenario 2. In this scenario, the professionals investigated EHRs in which the patients reported the absence of a fever, referred not to be allergic, and only had a cough. The combination of elements in this scenario is relevant because a fever may be a warning sign of a disease that needs to be investigated more quickly. It opens the investigation of other pathologies with viral, bacterial, and fungal origin. The cough not associated with an allergy leads the physician to the investigation of other parameters that might identify the cause of the cough. In addition, through the absence of a fever it is possible to avoid subjecting the patient to more severe treatments. This

scenario was composed of a more complex query that involved three sentences as follows:

1. **Keyword**: "Cough" $\wedge$ **Dimension**: $< Invention > \wedge$ **Value**: "Refers";
2. **Keyword**: "Allergy" $\wedge$ **Dimension**: $< Mode > \wedge$ **Value**: "Deny";
3. **Keyword**: "Fever" $\wedge$ **Dimension**: $< Mode > \wedge$ **Value**: "Deny";

4.2. Datasets

Annotation of illocutions. The pragmatic search mechanism is based on annotations of illocutions related to the content of EHRs. The fully automatic detection and annotation of illocutions in free-text messages in NL still remains a research challenge (Reis *et al*., 2017), for it required the involvement of participants to analyze the EHR contents so as to insert the annotations regarding the illocutions. The annotation process involved the reading of the EHRs, and the definition of illocution types and values related to them. For instance, given a description of an EHR in NL, the participant associated a fragment of text with CUIs (Concept Unique Identifier) existing in UMLS (*e.g.*, the concept of "Pain") to identify synonyms. The concept was related to an illocution dimension, and a value was given to such dimension. For example, the concept of "Pain" is related to the dimension of $Time$ and the value "30 days". The participants used features of the *SiRBI* system implemented to facilitate this process.

Base of reference. This stage aimed to classify the relevance of EHRs into a subset of the experimental dataset for each of the scenarios under study. The construction of a reference dataset is necessary because there is no standardized test set in the context of this research. Participants were invited to review the content of EHRs from a subset of the database. They needed to indicate the degree of relevance of each EHR to the objective of IR according to the scenarios' requirements. For the elaboration of the subset of EHRs (to avoid the analysis of the whole database with 13.300 EHRs), we considered the search results of the expanded search, which involves the search of EHRs with all possible variations of related keywords based on the expanded queries considering the UMLS (around 3 hundreds). According to the defined search scenarios, the health professionals classified EHRs among irrelevant, low relevant, relevant, and very relevant.

4.3. Configurations and Evaluation Metrics

In order to verify the results of the pragmatic search developed in the *SiRBI* system, each proposed search scenario was tested by exploring four configurations: (i) *Solr*[3]; (ii) Syntactic search with the *SiRBI*; (iii) Expanded search exploring the *UMLS*; (iv) and the Pragmatic search, as proposed in this work. Considering the constructed *Dataset* for the scenarios, our goal was to analyze the capacity of the pragmatic search

3. http://lucene.apache.org/solr/

elaborated in the *SiRBI* to improve, through the annotations and proposed algorithm, the results obtained in the syntactic and expanded search. Thus, we do not intend to provide an exhaustive competition between existing approaches. We are primarily interested in analyzing the contributions obtained between configurations (ii) and (iii) in relation to configuration (iv). The goal was to verify the benefits of the approach when we use the *PraSA* algorithm in conjunction with existing tools such as *Solr*.

Configuration (i) was introduced to comparatively illustrate the behavior of the *Solr* Search engine when directly implemented and used in the administration interface, because it presents additional filters and extensions to those used in configuration (ii). For this purpose, queries in configuration (i) were made by using the keywords from the respective scenario, including the dimension values, in addition to the terms used in configuration (ii). For instance, the query in configuration (i) for scenario 1 was formulated as follows ("Cough" $\wedge$ "At night" $\wedge$ "Deny" $\wedge$ "Allergy"), whereas the structured query in the configuration (ii) – Syntactic with the *SiRBI* – only included ("Cough" $\wedge$ "Allergy"). As defined in the *PraSA* algorithm, the expanded search – configuration (iii) – considers the semantic expansion of terms from the *UMLS* over the search results obtained in the Syntactical search – configuration (ii).

To assess the quality of the responses provided by the *SiRBI* system in the distinct configurations, we calculated the standard IR metrics of *Precision*, *Recall* and *F-Measure*, which are widely used in the field of IR. We computed *Precision* as the number of relevant and very relevant EHRs correctly retrieved (according to the specialists) by the search configuration in contrast to the constructed dataset, over the total number of EHRs retrieved by the search mechanism (result of the execution of each configuration). We calculated the *Recall* as the number of relevant and very relevant EHRs correctly retrieved over the total number of Relevant and Very Relevant EHRs that resulted from the expanded search configuration. The results of the expanded search mechanism were assumed as the total possible set of EHRs recovered (set of EHRs that was evaluated by the participants). Therefore, we considered it as the divisor to calculate the *Recall*. The *F-measure* refers to the harmonic mean of *Precision* and *Recall*.

4.4. Results

We present the results by organizing them into the two search scenarios investigated. Table 2 shows the results for scenario 1 (Patients who report a cough at night but do not present any type of allergy) in the following configurations: (i) *Solr*; (ii) Syntactic with the *SiRBI*; (iii) Expanded search using UMLS; (iv) and the Pragmatic search mechanism as proposed in this investigation.

Results indicate that the search configuration with the *Solr* presents a better *Precision* (0.67), but a lower *Recall* (0.43) as compared to the Syntactic configuration using the *SiRBI*. The configuration with *Solr*, which explored the keywords, dimensions, and values in the search query, reached a *F-Measure* of 0.52. It surpasses the results from the Syntactic and the Expanded configuration. The expanded search reaches a total

value for *Recall*, but significantly decreases *Precision* (0.12). Our pragmatic search proposal provides an overall better result, harmonizing *Precision* and *Recall*. The obtained *F-Measures* achieved 0.92.

Table 2. Results for Scenario 1

	Solr	**Syntactic**	**Expanded**	**Pragmatic** (*SiRBI*)
#Irrelevant	3	25	72	0
#Low Relevant	5	179	199	2
#Relevant	5	10	12	8
#Very Relevant	11	25	25	25
#TOTAL	24	239	308	35
Precision	0.67	0.15	0.12	0.94
Recall	0.43	0.95	1.0	0.89
F-Measure	0.52	0.25	0.21	0.92

Following the same configurations, Table 3 presents the results for the second search scenario (EHRs containing patients who report the absence of a fever, report that they are allergic, and only have a cough). Results point out the values of *Precision* and *Recall* for the *Solr* and the *Syntactic* configuration. The latter improved with respect to the scenario 1 (*cf.* Table 2). Similar to the findings in scenario 1, the *Solr* configuration overcomes the *Syntactic* configuration, but here with a small difference. The results for the expanded search also improved, but the *F-Measure* (0.58) was still lower than the *Solr* and the *Syntactic* configuration. The Pragmatic search presents the best results for scenario 2, with a *F-Measure* of 0.76. Although this result is less expressive than in scenario 1, it still presents a gain with respect to the other evaluated search configurations. The second scenario explored a more complex query compared to scenario 1 by involving additional combinations of concepts.

Table 3. Results for Scenario 2

	Solr	**Syntactic**	**Expanded**	**Pragmatic** (*SiRBI*)
#Irrelevant	45	53	90	3
#Low Relevant	51	54	65	19
#Relevant	44	42	49	34
#Very Relevant	54	54	59	45
#TOTAL	194	203	263	101
Precision	0.51	0.47	0.41	0.78
Recall	0.91	0.89	1.0	0.73
F-Measure	0.65	0.62	0.58	0.76

5. Discussion

The experimental results indicated the potential of the *SiRBI* system when considering the users' intentions in EHR-based investigative scenarios. The results showed a significant improvement provided by the Pragmatic Search when compared to the Expanded Search and Syntactic Search. The obtained results of the empirical experiments are certainly advances that demonstrate the usefulness of the proposed method, which considers Pragmatics search of EHRs.

In scenario 1, the tested query was specific in that it considered patients who report a "cough at night" and "deny allergy". The Syntactic search retrieved EHRs only containing the terms "cough" and "allergy", regardless of the parameters related to the dimension and value of illocutions. In contrast, the *Solr* configuration considered dimension values. The results of EHRs evaluated as "Irrelevant" and "Little Relevant" are due to the fact that Syntactic search implemented via the *SiRBI* and the *Solr* recovered EHRs that contained the terms "cough" and "allergy", but diverged in relation to the illocution dimension and its value in the query. For example, EHRs that refer to "cough" and present some kind of "allergy" are returned. The results of the EHRs classified as "Relevant" and "Very Relevant" present documents with keywords, dimensions, and values as requested in the query.

The results retrieved through the expanded Search configuration presented a high number of "Irrelevant" and "Low Relevant" EHRs (Table 2). This configuration has a lower *Precision* due to the fact that more combinations of domain terms were queried (*i.e*, it considered the combination of all terms related to "cough" and "allergy"). This is particularly important when the person who searches the data has a different vocabulary from the person who produced the content (which is frequent in multidisciplinary groups), but it decreases *Precision*. The Pragmatic Search showed good results for the Scenario 1 by eliminating EHRs that do not accurately express the query parameters, especially EHRs not related to the intentions of the person who is searching. Consequently, there was an improvement in the result of *Precision* and *F-Measure*.

Scenario 2 proposed a more complex query, although terms like "cough", "allergy" and "fever" are easier to find in EHRs when compared to a more specific term such as "nocturnal cough". The evaluations presented a higher number of "Irrelevant" and "Low Relevant" EHRs for both Syntactic search with *SiRBI* and the *Solr* configuration. We observed a similar result with the Expanded search configuration. The expanded search attempted to retrieve EHRs related to the terms expressed in the query. This resulted in the increase of 60 results of which only 12 were assigned as "Relevant". It revealed a low *Precision* when we make the Cartesian product among the related terms. As in scenario 1, the Pragmatic Search demonstrated an improvement in quality, since it considered the relationship between intentions and concepts in EHRs annotations. It performed a more consistent filter in the search results according to the available metadata. It eliminated EHRs that were not related to the illocution dimensions and values, providing a more adequate selection of the search results.

Future work involves the investigation of a semi-automatic method for the annotation of illocutions in NL texts. Although our proposal has been considered computationally and operationally feasible (for "small" databases), the study of an automatic annotator is desirable for use in large-scale databases. An automatic annotation can enable practical long-term evaluations, instead of evaluations based on representative scenarios. The manual analysis of the annotations allowed for the identification of some patterns that we have already studied and experimented (Reis *et al*., 2017). Our goal is to develop an annotation mechanism using domain ontologies for the detection of illocutions, and, consequently, improve search quality. For example, an annotation

regarding "cough at night" is equivalent to "nocturnal cough". A refinement of annotations in this sense can enhance the quality of the results in the Pragmatic Search. We also aim to further investigate the effects of ranking results.

Another aspect to be considered is the integration of the *SiRBI* system with collaborative platforms. This is a key step towards the "large and long-term" practical use and analysis of the influence of considering aspects related to users' intentions in IR, and sharing of collective intelligence. In addition, other models for representing additional intention related elements can be explored to improve our approach.

6. Conclusion

The use of text-based fields in EHRs remains an important mechanism for clinical research and communication among health care professionals. The ability to recognize the others' intentions remains a relevant element for the social sensibility of a group. Physicians express their intentions on text-based content in EHRs. However, poorly structured texts described in NL make it difficult to retrieve and share these data adequately. This article proposed a framework that considers intention aspects to benefit IR and sharing by multidisciplinary groups. This research contributed with an IR method based on illocutions, and defined an algorithm to filter and select the EHRs. The proposed concepts were thoroughly implemented in the *SiRBI* system. We presented an experimental evaluation to assess the proposed IR method. The conducted experiment allowed us to ascertain the benefits of the proposal through an analysis of the distinct configurations using standard measures. Our experimental tests demonstrated the overall effectiveness of the retrieval technique developed in this investigation, and its benefits to the medical field. In particular, this technique has shown to be effective when considering situations in which various health care professionals should share detailed patients' data. The Pragmatic Search obtained the best *F-Measure* values in the evaluated scenarios, with expressive advances in the *Precision* of the search results when compared to the Syntactic Search and Expanded Search. Future work involves additional investigations and experiments to refine the results. We aim to improve the annotation process, and to analyze long-term usage findings by considering further parameters of the *SiRBI*.

Acknowledgements

This work is supported by the São Paulo Research Foundation (FAPESP) (Grant #2017/02325-5)[4].

References

Alyami M. A., Almotairi M., Aikins L., Yataco A. R., Song Y. T. (2017, May). Managing personal health records using meta-data and cloud storage. In *2017 ieee/acis 16th international conference on computer and information science (icis)*, p. 265-271.

Apperly I. (2010). *Mindreaders: The cognitive basis of "theory of mind"*. Taylor & Francis.

Aroma R. J., Kurian M. (2013, March). A semantic web: Intelligence in information retrieval. In *2013 ieee international conference on emerging trends in computing, communication and nanotechnology (iceccn)*, p. 203-206.

Asai H., Yamana H. (2014). Intelligent ink annotation framework that uses user's intention in electronic document annotation. In *Proceedings of the ninth acm international conference on interactive tabletops and surfaces*, p. 333–338. New York, NY, USA, ACM.

Austin J. (1962). *How to do things with words?* Clarendon Press, London.

Bhatt M., Rahayu W., Soni S. P., Wouters C. (2009). Ontology driven semantic profiling and retrieval in medical information systems. *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 7, No. 4, p. 317 - 331. (Semantic Web challenge 2008)

Bodenreider O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, Vol. 32, No. suppl_1, p. D267-D270.

Bonacin R., Reis J. C. D., Hornung H., Baranauskas M. C. C. (2012). An ontological model for representing pragmatic aspects of collaborative problem solving. In *2012 ieee 21st international workshop on enabling technologies: Infrastructure for collaborative enterprises*, p. 444-449.

Chawla S., Bedi P. (2008, Aug). Query expansion using information scent. In *2008 international symposium on information technology*, Vol. 3, p. 1-8.

Chen T.-L., Chung Y.-F., Lin F. Y. (2012, June). A study on agent-based secure scheme for electronic medical record system. *J. Med. Syst.*, Vol. 36, No. 3, p. 1345–1357.

Cogley J., Stokes N., Carthy J. (2013). Exploring the effectiveness of medical entity recognition for clinical information retrieval. In *Proceedings of the 7th international workshop on data and text mining in biomedical informatics*, p. 3–4. New York, NY, USA, ACM.

Dong H., Hussain F. K., Chang E. (2008, Feb). A survey in semantic search technologies. In *2008 2nd ieee international conference on digital ecosystems and technologies*, p. 403-408.

Forcher B., Roth-Berghofer T., Agne S., Dengel A. (2014). Intuitive justifications of medical semantic search results. *Engineering Applications of Artificial Intelligence*, Vol. 30, p. 1 - 17.

Gómez M. Montes-y, Gelbukh A. F., López-López A. (1999). Document title patterns in information retrieval. In V. Matousek, P. Mautner, J. Ocelíková, P. Sojka (Eds.), *Text, speech and dialogue: Second international workshop, tsd'99 plzen, czech republic, september 13–17, 1999 proceedings*, p. 372–375. Berlin, Heidelberg, Springer Berlin Heidelberg.

Guha R., McCool R., Miller E. (2003). Semantic search. In *Proceedings of the 12th international conference on world wide web*, p. 700–709. New York, NY, USA, ACM.

Gupta V., Garg N., Gupta T. (2012, Feb). Search bot: Search intention based filtering using decision tree based technique. In *2012 third international conference on intelligent systems modelling and simulation*, p. 49-54.

Gurulingappa H., Müller B., Hofmann-Apitius M., Fluck J. (2011). A semantic platform for information retrieval from e-health records. In *Proceedings of the twentieth text retrieval conference, TREC 2011, gaithersburg, maryland, usa, november 15-18, 2011.*

Hesse B. W., Connell M. O., Augustson E. M., Chou W.-Y. S., Shaikh A. R., Rutten L. J. F. (2011, jul). Realizing the promise of web 2.0: Engaging community intelligence. *Journal of Health Communication*, Vol. 16, No. sup1, p. 10–31.

Hesse B. W., Hansen D., Finholt T., Munson S., Kellogg W., Thomas J. C. (2010, Nov). Social participation in health 2.0. *Computer*, Vol. 43, No. 11, p. 45-52.

Hildebrand M., Ossenbruggen J. V., Hardman L. (2007). *An analysis of search-based user interaction on the semantic web.*

Hwang M., Kim P., Choi D. (2011, June). Information retrieval techniques to grasp user intention in pervasive computing environment. In *2011 fifth international conference on innovative mobile and internet services in ubiquitous computing*, p. 186-191.

Khan A. S., Fleischauer A., Casani J., Groseclose S. L. (2010). The next public health revolution: Public health information fusion and social networks. *American Journal of Public Health*, Vol. 100, No. 7, p. 1237-1242.

Koopman B., Zuccon G., Bruza P., Sitbon L., Lawley M. (2016, Apr 01). Information retrieval as semantic inference: a graph inference model applied to medical search. *Information Retrieval Journal*, Vol. 19, No. 1, p. 6–37.

Kreuzthaler M., Daumke P., Schulz S. (2015). Semantic retrieval and navigation in clinical document collections. *Studies in health technology and informatics*, Vol. 212, p. 9-14.

Laforest F., Tchounikine A. (1999). Indexing semi-structured documents for context-based information retrieval in a medical information system. In *Proceedings. tenth international workshop on database and expert systems applications. dexa 99*, p. 593–597.

Lévy P. (2001). *Cyberculture*. University of Minnesota Press.

Liaw S.-T., Taggart J., Yu H., Lusignan S. de, Kuziemsky C., Hayen A. (2014). Integrating electronic health record information to support integrated care: Practical application of ontologies to improve the accuracy of diabetes disease registers. *Journal of Biomedical Informatics*, Vol. 52, p. 364 - 372.

Liu K. (2000). *Semiotics in information systems engineering*. Cambridge University Press.

Liu K., Li W. (2014). *Organisational semiotics for business informatics*. Taylor & Francis.

Mendoza M., Baeza-Yates R. (2008, Oct). A web search analysis considering the intention behind queries. In *2008 latin american web conference*, p. 66-74.

Morris C. (1937). *Logical positivism, pragmatism and scientific empiricism*. Hermann et cie.

Noor S., Martinez K. (2009). Using social data as context for making recommendations: An ontology based approach. In *Proceedings of the 1st workshop on context, information and ontologies*, p. 7:1–7:8. New York, NY, USA, ACM.

Peirce C. S. (1931–1935). *Collected papers*. Cambridge, Harvard Universit Press.

Perez-Rey D., Jimenez-Castellanos A., Garcia-Remesal M., Crespo J., Maojo V. (2012, Apr 05). Cdapubmed: a browser extension to retrieve ehr-based biomedical literature. *BMC Medical Informatics and Decision Making*, Vol. 12, No. 1, p. 29.

Popescu M. (2010, July). An ontological fuzzy smith-waterman with applications to patient retrieval in electronic medical records. In *International conference on fuzzy systems*, p. 1-6.

Pruski C., Wisniewski F. (2012). Efficient medical information retrieval in encrypted electronic health records. *Studies in health technology and informatics*, Vol. 180, p. 225-9.

Reis J. C. dos, Bonacin R., Baranauskas M. C. C. (2017). Recognizing intentions in free text messages: Studies with portuguese language. In *Ieee 26th international conference on enabling technologies: Infrastructure for collaborative enterprises*, p. 302-307.

Reis J. C. dos, Bonacin R., Perciani E. M. (2016a). Intention-based information retrieval of electronic health records. In *Ieee 25th international conference on enabling technologies: Infrastructure for collaborative enterprises (wetice)*, p. 217-222.

Reis J. C. dos, Bonacin R., Perciani E. M., Baranauskas M. C. C. (2016b). Analysis and representation of illocutions from electronic health records. In M. C. C. Baranauskas, K. Liu, L. Sun, V. P. d. A. Neris, R. Bonacin, K. Nakata (Eds.), *Socially aware organisations and technologies.*, p. 209–218. Cham, Springer International Publishing.

Searle J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge, London, Cambridge University Press.

Sharef N. M., Madzin H. (2012, March). Ims: An improved medical retrieval model via medical-context aware query expansion and comprehensive ranking. In *2012 international conference on information retrieval knowledge management*, p. 214-218.

Sweet L. E., Moulaison H. L. (2013, Dec 01). Electronic health records data and metadata: Challenges for big data in the united states. *Big Data*, Vol. 1, No. 4, p. 245-251. Retrieved from https://doi.org/10.1089/big.2013.0023

Tang X., Liu K., Cui J., Wen F., Wang X. (2012, July). Intentsearch: Capturing user intention for one-click internet image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34, No. 7, p. 1342-1353.

Tawfik A. A., Kochendorfer K. M., Saparova D., Ghenaimi S. A., Moore J. L. (2011, June). Using semantic search to reduce cognitive load in an electronic health record. In *Ieee 13th international conference on e-health networking, applications and services*, p. 181-184.

Totelin L. (2009). *Hippocratic recipes: Oral and written transmission of pharmacological knowledge in fifth- and fourth-century greece*. Brill.

Zheng K., Mei Q., Hanauer D. A. (2011). Collaborative search in electronic health records. *Journal of the American Medical Informatics Association*, Vol. 18, No. 3, p. 282-291.

Zinglé H. (2006). Modelling knowledge with zdoc for the purposes of information retrieval. In M. Ali, R. Dapoigny (Eds.), *Proceedings of 19th international conference on industrial, engineering and other applications of applied intelligent systems, iea/aie 2006, annecy, france, june 27-30, 2006. proceedings*, p. 1053-1058. Springer Berlin Heidelberg.