Expert Systems With Applications 44 (2016) 386–399

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: An UMLS approach

Israel Alonso*, David Contreras

Department of Telematics and Computer Science, Comillas Pontifical University, C/ Alberto Aguilera, 25, 28015 Madrid, Spain

ARTICLE INFO

Keywords:
Semantic similarity
Information retrieval
Electronic Health Record
UMLS

ABSTRACT

One promise of current information retrieval systems is the capability to identify risk groups for certain diseases and pathologies based on the automatic analysis of vast amounts of Electronic Medical Records repositories. However, the complexity and the degree of specialization of the language used by the experts in this context, make this task both challenging and complex. In this work, we introduce a novel experimental study to evaluate the performance of the two semantic similarity metrics (*Path* and *Intrinsic IC-Path*, both widely accepted in the literature) in a real-life information retrieval situation. In order to achieve this goal and due to the lack of methodologies for this context in the literature, we propose a straightforward information retrieval system for the biomedical field based on the UMLS Metathesaurus and on semantic similarity metrics. In contrast with previous studies which focus on testbeds with limited and controlled sets of concepts, we use a large amount of information (101,712 medical documents extracted from TREC Medical Records Track 2011). Our results show that in real-life cases, both metrics display similar performance, *Path* (F-Measure = 0.430) e *Intrinsic IC-Path* (F-Measure = 0.427). Thereby we suggest that the use of *Intrinsic IC-Path* is not justified in real scenarios.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

The exponential growth, in recent times, of the amount of biomedical information that is stored on purely electronic supports — Electronic Health Records, or EHR, spring promptly to our mind — has turned them into an element of undeniable relevance to the most diverse fields of scientific research (Hoffman, 2010; Prokosch, & Ganslandt, 2009).

One of these fields is that of Information Retrieval, and its traditional challenge of identifying those records which most efficiently answer a user's immediate needs for information; for this task to be accomplished, it is critical to first establish a recognition of patterns in medical histories which would permit, ultimately, the early detection of epidemic outbreaks, the prevention of disease, or the identification of cohort groups (Roque, et al., 2011). The main difficulty in undertaking this task arises from Natural Language Processing, as natural language is not only complex, but also highly context-sensitive. In a broad field such as that of the English language, for instance, it becomes necessary draw upon resources and ontologies like WordNet to aid representation (Fellbaum, 1998). Unfortunately, these tools are of limited use to more specialized disciplines, such as that of biomedicine, whose technical jargon is often as complex as it is ambiguous; the parsing of biomedical information calls for very specific terminology (Friedman, Kra, & Rzhetsky, 2002) and, hence, for new search strategies, designed from the outset to the particular demands of this branch of science (Alpi, 2005). In such cases, one must resort to specialist resources — dictionaries and thesauri like UMLS (McCray et al., 1993) — to give a semantic value to relevant information.

Our present work aims to bridge this gap, helping the information retrieval systems based on Electronic Health Records, according to their semantic content; in a nutshell, being able to interpret the information needs of any given query, and consequently select those medical documents most relevant in terms of semantic proximity. An endeavor which is, we believe, much needed for the correct identification of patients in cohort studies, given the complexity, variability, and lack of structure in the information traditionally contained in such records. This will require, to define and represent, through biomedical concepts, the information contained in both health records and medical queries, in order to establish the semantic proximity between them. The use, in this fashion, of semantic relationships between said concepts, closely emulates the analogous process in the human mind to establish similarity between two given terms (Miller, & Charles, 1991; Rubenstein, & Goodenough, 1965). It should be pointed out beforehand that previous works have

* Corresponding author. Tel.: +34 915422800; fax: +34 91 559 65 69.
E-mail addresses: ialonso@comillas.edu (I. Alonso), davidcb@comillas.edu (D. Contreras).

http://dx.doi.org/10.1016/j.eswa.2015.09.028
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

shown interest in establishing metrics for determining the degree of semantic similarity between two terms (Collins, & Loftus, 1975) — in a more general context like the English language, and based on the WordNet infrastructure (Meng, Huang, & Gu, 2013). Unfortunately, however, these approaches not always yield satisfactory results when they are applied in the biomedical domain, since WordNet's coverage of this domain is rather limited. (Burgun, & Bodenreider, 2001). Later works have attempted to solve this by incorporating specific resources and ontologies (MeSH, and SNOMED CT) in the study of similarity metrics in the field of biomedicine, always in a theoretical context and a controlled environment. (Al-Mubaid & Nguyen, 2006; Batet, Sánchez, & Valls, 2011; Caviedes, & Cimino, 2004; Nguyen, & Al-Mubaid, 2006; Pedersen, Pakhomov, Patwardhan, & Chute, 2007). These works prove it becomes necessary to resort to a specialised infrastructure — namely UMLS — if we are to determine the similarity existing between two concepts in the field of biomedicine, with the degree of precision that a human expert would expect to achieve.

In this work we propose an experimental study to evaluate the performance of the two semantic similarity metrics (*Path* and *Intrinsic IC-Path*, both widely accepted in the literature) in a real-life information retrieval context. Moreover, to perform this assessment, we deploy a straightforward information retrieval system for the biomedical field based on the UMLS Metathesaurus and on semantic similarity metrics, due to the lack of methodologies for this context in the literature.

Our paper will be structured as follows:

In Section 2, we will describe the main components and characteristics of UMLS. In Section 3, we offer an outline of the current state of the art, focusing on different tools and strategies used nowadays in the retrieval of biomedical information, as well as the metrics used in calculating the semantic similarity between two concepts in this particular field. Then, in Section 4, we will define our proposal, along with the materials used in our work. In Section 5, we conduct a study of the inner workings of the different sources and relationships contained in UMLS, and how they are reflected in the results obtained by semantic similarity metrics in a purely theoretical context; we will later use, as our reference, the two main metrics based *on* the approaches *Intrinsic IC* and *Path finding* for their study and their application to a real-life context; Section 6 will describe the procedures involved in our proposal for an ad-hoc and straightforward concept-based medical document retrieval system, and evaluate the efficacy of the two main semantic similarity metrics when applied to a real-life context (reflected in *TREC 2011*). Section 7 covers the analysis and interpretation of the results obtained. Last, Section 8 will comment on the conclusions derived from all conducted tests, as well as the contributions obtained from their results, and the future lines of research that would give continuity to our work.

2. UMLS

UMLS[1] (Unified Medical Language System) is an ongoing project started in 1986 by the National Library of Medicine. It was envisioned as a common environment for the access and treatment of biomedical information (Bodenreider, 2004; Humphreys, Lindberg, Schoolman, & Barnett, 1998; Lindberg, Humphreys, & McCray, 1993). To this end, it structures said information as a series of concepts, with a set relationship between them. At its core, UMLS is made up of three components, all of which undergo regular updates and revision: a Metathesaurus, a Semantic Network, and a Specialist Lexicon (lexical information and tools for natural language processing). Of these elements, the Metathesaurus and the Semantic Network are of particular interest to our work: the former for its contained concepts, sources and relationships, and the latter for its offer of semantic types.

Table 1
Representation structure of UMLS concept C0018787.

CUI	LUI	SUI	AUI	Source	String
C0018787	L0018787	S0047194	A0066368	MeSH	Heart
C0018787	L0018787	S0047194	A16757661	NCI	Heart
C0018787	L0018787	S0047194	A2882201	SNOMED	Heart
C0018787	L0018787	S0375948	A16766657	NCI	HEART
C0018787	L0018787	S0419735	A0480532	CSP	heart
C0018787	L0018787	S0419735	A18628913	CHV	heart
C0018787	L0248647	S0324326	A12802806	NCI	Cardiac
C0018787	L0248647	S1344787	A1304355	CSP	Cardiac
C0018787	L0248647	S1344787	A18647556	CHV	Cardiac

The Metathesaurus is, in essence, a vast multipurpose and multi-language database covering more than one million concepts, all of them represented under a common framework, and stored in over a hundred different sources. Said sources are grouped in several distinct perspectives of the biomedical environment, such as scientific information (MeSH-CRISP), clinical terminology (SNOMED-CT), administrative terminology (ICD-9-CM, CPT-4), or data exchange (HL7, LOINC), as well as general or specific thesauri including anatomy (UWDA, NeuroNames), drugs (RxNorm, First Data Bank), medical devices (UMD, SPN), nursing (NIC, NOC, NANDA), oncology (PDG), adverse reactions (COSTART, WHO) or gene products (Gene Ontology-GO), to name a few.

The data compiled in these various sources is organized in the Metathesaurus following a unique identifier structure, with a hierarchy of four significance levels: Concepts, Terms, Strings, and Atoms. In this order:

- CUI (Concept Unique Identifier): Each concept represents a distinct meaning, which encompasses, within a unique code, all its synonym terms.
- LUI (Lexical Unique Identifier): Identifies each of the known lexical variations or terms for any given concept.
- SUI (String Unique Identifier): Represents each descriptive string associated to a given term. One of them is designated as its name, or preferred term. All predicted variations in the character sequence of the string (upper and lower case, punctuation) are covered in separate identifiers.
- AUI (Atom Unique Identifier): correspond to each individual occurrence of a given string in a specific source.

Hence, for instance, the concept (C0018787), which represents the muscle organ that keeps blood circulation going, is grouped into a number of descriptive strings, of which we now show a few for the sake of the example. (Table 1).

We must keep in mind that a given descriptive string (SUI), may be referenced in one or many concept identifiers (CUIs). For example, the string "*Heart*", identifies the preferred term for concept (C0018787), but it is also one of the synonym terms for concept (C1281570) "*Entire heart*". We will now show the series of descriptive strings for both these concepts, as well as the semantic type they belong to:

CUI: C0018787
SUI (Prefered term): Heart
Other SUIs (string terms): Hearts; Cardiac; coronary; cardiac structure; heart structure; structure of heart, unspecified; corazón; estructura cardiaca; Cuore; herzen; Hart; etc.
Semantic Type: (bpoc) - Body Part, Organ, or Organ Component.

CUI: C1281570
SUI (Prefered term): Entire heart
Other SUIs (string terms): Heart; Entire heart (body structure); corazón; etc.
Semantic Type: (bpoc) - Body Part, Organ, or Organ Component.

[1] http://www.nlm.nih.gov/research/umls/.

Download English Version:

https://daneshyari.com/en/article/382050

Download Persian Version:

https://daneshyari.com/article/382050

Daneshyari.com