A Lexical-Ontological Resource for Consumer Heathcare

Elena Cardillo

FBK-IRST, Via Sommarive 18, 38123 Trento, Italy
cardillo@fbk.eu

Abstract. In Consumer Healthcare Informatics it is still difficult for laypersons to understand and act on health information, due to the persistent communication gap between specialized medical terminology and that used by healthcare consumers. Furthermore, existing clinically-oriented terminologies cannot provide sufficient support when integrated into consumer-oriented applications, so there is a need to create consumer-friendly terminologies reflecting the different ways healthcare consumers express and think about health topics. Following this direction, this work suggests a way to support the design of an ontology-based system that mitigates this gap, using knowledge engineering and Semantic Web technologies. The system is based on the development of a consumer-oriented medical terminology which will be integrated with other existing domain ontologies/terminologies into a medical ontology repository. This will support consumer-oriented healthcare systems by providing many knowledge services to help users in accessing and managing their healthcare data.

Keywords: Medical Knowledge Acquisition, Knowledge Integration, Medical Ontologies, Consumer Medical Terminologies.

1 Introduction

With the advent of the Social Web and Healthcare Informatics technologies, we can recognize that a linguistic and semantic discrepancy still exists between specialized medical terminology used by healthcare providers or professionals, and the so called “lay” medical terminology used by healthcare consumers. The medical communication gap became more evident when consumers started to play an active role in healthcare information access, becoming more responsible for their personal healthcare, exploring health-related information sources on their own, consulting decision-support healthcare sites on the web, and using patient-oriented healthcare systems which allow them to directly read and interpret clinical notes or test results and to fill in their Personal Health Record (PHR). To help consumers fill this gap, the challenge is to sort out the different ways they communicate within distinct discourse groups and map the common, shared expressions and contexts to the more constrained, specialized language of healthcare professionals. In particular, medical Knowledge Integration in healthcare systems is facilitated by the use of Semantic Web technologies,

A. Bernstein et al. (Eds.): ISWC 2009, LNCS 5823, pp. 949–956, 2009.

helping consumers during their access to healthcare information and improving the exchange of their personal clinical data. Though much effort has been spent on the creation of these medical resources, used above all to help physicians in filling in Electronic Health Records (EHR), there is little work based on the use of consumer-oriented medical terminology, and in addition most existing studies have been done only for English.

Given this scenario, this work want to propose a methodology for the creation of a consumer-oriented lexical-ontological resource for Italian, and its integration with a coherent semantic medical resource, which could be used in healthcare systems, like Personal Health Records, to help consumers during the process of querying and accessing healthcare information. The present work will be structured as follows: In Section 2 is described the State of the Art in the field of medical terminologies/ontologies, both in consumer-oriented and clinically-oriented healthcare; in Section 3 are exposed the problem statement, our objectives and approach to reach them; in Section 4 are presented preliminary results; and finally in Section 5 are proposed concluding remarks and some future works.

2 State of the Art

2.1 Medical Terminologies and Ontologies

Over the last two decades the standardization efforts have established a number of medical terminologies and classification systems as well as conversion mappings between them to help medical professionals in managing and codifying their patients health care data, such as UMLS Metathesaurus, SNOMED, ICD-10 (International Classification of Diseases) and the ICPC-2 (International Classification of Primary Care). They concern *"the meaning, expression, and use of concepts in statements in the medical records or other clinical information systems"* [11]. In the presence of all these medical terminologies interoperability has become a significant problem. Content, structure, completeness, detail, cross-mapping, taxonomy, and definitions vary between existing vocabularies. So during the last few years, thanks to the Semantic Web perspective, the collaboration between the areas of Healthcare Informatics and Knowledge Representation generated a set of new methodologies and tools for improving healthcare systems, and in particular medical terminologies, which were translated into more formal representations using ontology languages (e.g. the logical formalization of SNOMED CT).

During the last few years much effort has also been spent on the creation of new Biomedical Ontologies (e.g. the Foundational Model Anatomy -FMA- [9]). Ontologies become relevant in healthcare if integrated into an EHRs, which manage an increasing volume of narrative data, to allow: structuring and semantics of the recorded information; and references to concepts from ontologies of the first kind, e.g. ICD 10/9 or SNOMED terms [5]. Two other important issues to take into account, given the presence of all these medical ontologies are: Ontology Mapping, to show how concepts of one ontology are semantically related to concepts of another ontology [6]; and Ontology Integration, which allows access to multiple heterogeneous ontologies. Much work has been done in this direction, for the alignment of different Biomedical Ontologies with concept overlap (here can be mentioned the works of Mork and Bernstein [8], and Zhang and Bodenreider [13]), and for their integration by means of medical ontology repositories, such as the creation of Bio-Portal, a Web-based system that serves as a repository for biomedical ontologies [10].

2.2 Consumer-Oriented Medical Vocabularies

In spite of these advantages reached by the integration of Healthcare Informatics and Semantic Web technologies, the vocabulary problem continues to plague health professionals and their information systems, and in particular laypersons who are the most damaged by the increased medical linguistic gap. To respond this healthcare consumers' needs, during the last few years, many researchers have labored over the creation of lexical resources that reflect the way consumers/patients express and think about health topics. One of the largest initiatives in this direction is the Consumer Health Vocabulary Initiative[1], resulted in the creation of the Open Access Collaborative Consumer Health Vocabulary (OAC CHV) for English. It includes lay medical terms and synonyms connected to their corresponding technical concepts in the UMLS Metathesaurus. They combined corpus-based text analysis with a human review approach, including the identification of consumer forms for "standard" health-related concepts. An overview of all these studies can be found elsewhere [7].

It is important to stress that there are only few examples of the real application of the most of initiatives. For example, in Zeng *et al.* [14] there is an attempt to face syntactic and semantic issues in the effort to improve PHRs readability, using the CHV to map content in EHRs and PHRs. On the other hand, Rosembloom *et al.* [12] developed a clinical interface terminology, a systematic collection of healthcare-related phrases (terms) to support clinicians' entries of patient-related information into computer programs such as clinical "note capture" and decision support tools, facilitating display of computer-stored patient information to clinician-users as simple human-readable texts.

3 Research Objectives and Directions

3.1 The Problem Statement

As mentioned in Section 1, healthcare consumers actually play an active role in accessing and managing their personal health care data. For this reason they need an easy and understandable access to medical information when using such healthcare systems, and at the same time physicians, from their side, need to understand patients reports on their conditions (severity pain, degree of discomfort). So communication terminology and understanding of medical concepts are

[1] http://www.consumerhealthvocab.org

serious barriers. This linguistic gap, in addition, also prevent full participation of consumers for example in shared health records, and often interferes in communication between patients and their health care providers. Furthermore, most of the existing "standards" medical terminologies and ontologies have been developed from the point of view of physicians, so they don't provide a sufficient support for their integration in all that applications designed for laypersons. This highlights the necessity of an intermediate consumer understandable terminology to be integrated with standard specific terminologies/ontologies in order to support the integration of consumer-oriented applications with that designed for experts. This thesis work will also focus on the nature of this medical communication gap in the Italian context, where there is a lack of consumer-oriented medical terminologies (as seen in Section 2, all previous works have been done for English), and where illiteracy, regional diversity, and the high presence of non-native speakers further intensify the problem.

3.2 The Objectives

The purpose of this work is to support the design of an ontology-based system that mitigates the language barrier between the healthcare consumer and professional medical domains. Knowing the forms used by laypersons and how such forms map to medical concepts is useful in assisting healthcare consumers to formulate queries and to understand retrieved medical documents, and also helps professionals and information systems to deal with patient inputs. The general aim can be divided into the following sub-objectives:

1. Development of a Consumer Medical Vocabulary for Italian, able to reflect the different ways consumers and patients express and think about health topics.
2. Integration of this "lay" terminology with other existing terminologies, in particular with the clinical ones relevant to reconstruct the process of care in General Practice.
3. Formal Representation in OWL language of these terminologies, and integration of them into a unique Medical Ontology Repository.
4. Implementation of Reasoning and Search services to support the development of semantic-based healthcare systems which need interchanges with patients and consumers.

3.3 Approach

The global approach followed for this research activity is divided in two macro phases. The first one includes the creation of a Consumer Health Vocabulary for Italian, for collecting common medical expressions and terms used by Italian speaking people. The second one focuses on the formal representation of medical terminologies which will be integrated with the developed consumers vocabulary, and the development of a Medical Ontology Repository in which all these ontologies and terminologies will be integrated. The activity will be characterized by the following tasks:

- Knowledge Acquisition/Terminology Extraction. Use of elicitation techniques to acquire all the lay terms, words, and expressions, used by laypeople to indicate specific medical concepts;
- Generation of the Italian Consumer Health Terminology. Selection of all the lay terms extracted that have been identified as good representatives of technical medical concepts, and consequent mapping analysis to a standard medical terminology.
- Formalization in terms of OWL. Medical terminologies such as ICD10 and ICPC2 will be formalized into OWL ontologies, and then will be integrated with the consumer-oriented medical vocabulary and other existing medical ontologies to guarantee semantic interoperability.
- Creation of a Medical Ontology Repository (MORe) and implementation of Knowledge Services. Some relevant resources will be integrated into MORe, an ontology collection that will be extended with a set of basic reasoning services to support the implementation of semantic based patient healthcare applications.

4 What Has Been Done So Far

4.1 Knowledge Acquisition Task

This first task aims at the acquisition of consumer-oriented terminology and knowledge about a specific subset of healthcare domain, and at the creation of the consumer-oriented medical vocabulary for Italian. A hybrid methodology was used for the identification of "lay" terms and expressions used by Italian speaking people to indicate "symptoms", "diseases", and "anatomical concepts". Three different target groups were considered: First Aid patients subjected to the Triage process; a community of high educated and middle age people; and finally a group of elderly people. In this methodology three different Elicitation Techniques were applied to the mentioned groups of people: 1) Collaborative wiki-based medical knowledge acquisition; 2) Nurse-assisted medical knowledge acquisition; and 3) Interactive medical knowledge acquisition combining traditional elicitation techniques (Focus Groups, Concepts Sorting and Games).

All the aquired knowledge was analysed by means of a term extraction tool (Text2Knowledge - T2K), which allowed to automatically extract terminology and to perform typical text processing techniques and statistical analyses (more details about the tool can be found in [1]). Term extracted were reviewed by two physicians to find incongruities done by laypeople in categorization of medical terms and in synonymy relations. Physicians have been also asked to map a term/medical concept pair by using a professional health classification system, the above mentioned International Classification for Primary Care 2nd Edition (ICPC2-E), which is used in particular by general practitioners for encoding symptoms, medical procedures and diagnosis. A more detailed description of the methodology for knowledge acquisition and of the term extraction process and mapping analysis can be found in Cardillo *et al.* [4].

First Results Evaluation. A variegated consumer-oriented terminology was acquired. From 225 Wiki pages 962 medical terms were extracted, and in particular were found 173 *Exact Mappings*, 80 *Related Mappings*, 94 *Hyperonyms*, 51 *Hypomyms* and, finally, 186 *Not Mapped* ICPC2 concepts. Most of the exact mappings to ICPC2 are related to anatomical concepts, and many synonyms were found for symptoms. Concerning the Nurse-assisted data set, from 2.000 Triage records 1108 relevant terms were extracted, providing mapping only for 726 terms. Here can be highlighted the high presence of lay terms used for expressing symptoms with exact mappings to ICPC2 (134 on a total of 240 exact mappings), but also many synonyms in lay terminology for ICPC2 concepts (386 *Related Mappings*). Finally, 321 medical terms were extracted by the Focus Group data set. Here all the symptoms extracted (79 terms) had corresponding medical concept in ICPC2 terminology (35 *Exact Mappings* and 44 *Related Mappings*).

The most profitable methodology for acquiring consumer-oriented medical terminology resulted the one assisted by Nurses. While Wiki-based method, even if not exploited for the collaborative characteristic, has demonstrated good qualitative and quantitative results. Comparing the three sets, the overlap is only of 60 relevant consumer medical terms. The overlap with ICPC2 is about 508 medical concepts on a total of 706 ICPC2 concepts. This means that all the other mapped terms can be considered synonyms or quasi synonyms of the ICPC2 concepts. The large number of not mapped terms and the low overlap between the three sets of extracted terms demonstrate that it was possible to extract a very variegated range of medical terms, many compound terms and expressions, which can be representative of the corresponding technical terms present in standard terminology, and which can be used as candidate for the construction of our consumer-oriented medical terminology for Italian.

4.2 OWL Encoding of Medical Classification Systems

A parallel activity to that of consumer-oriented terminology acquisition was performed to formalize two Medical Classification Systems into OWL ontologies [2]: the previously mentioned ICPC2 and ICD10, expressing the two ontologies according to the sublanguage DL (Description Logic). In the process of conversion of ICPC2-ICD10 to OWL formalism two important principles of classification have been preserved: the disjointness of terms (nodes) and the exhaustiveness of classification, by introducing the use of special groups of terms such as "other", "unspecified" and "not elsewhere classified", reflecting this property in OWL by the explicit definition of sibling classes as disjoint and by the closure definition of any subdivision class as to be equivalent to a disjunction of all its child classes (including other, unspecified and so on). In encoding ICD10 to OWL, every ICD10 chapter is a class and each section is a subclass, which contain in turn each three or four digit ICD items. So only the subsumption and the disjunction relations are defined, the concepts representing each ICD category are labelled by the ICD codes. In encoding ICPC2 we preserved its biaxal structure creating a class for each chapter (body system or problem area) and a class for each component (*Symptom and Complaint*; *Procedure*, and *Diagnosis and Disease*). We added disjoint statements between siblings and some objects and datatype properties (description, terms of inclusion, terms of exclusion, ICD10 corrispondence)[2].

A well-founded and medically sound mapping model between the two ontologies was constructed as well, by means of its formalization in terms of OWL axioms (686 in total) and the validation of its coherence using Semantic Web techniques. Standardly, given two heterogeneous representations, a mapping can be viewed as a triple $\langle e, e', r \rangle$, where e, e' are the entities (e.g., formula, terms, classes, etc.) belonging two the different representations, and r is the relation asserted by the mapping. Due to the idea of encoding ICPC-ICD mappings as OWL axioms, the entities in the mapping correspond to ICPC-2 and ICD-10 classes and expressions, while the relation r is given a set-theoretic meaning by using subsumption and equivalence. Details about methodology for formalization and results can be found in [3]. After the task of mapping analysis and the evaluation of the first results, the extracted "lay" terms considered as good synonyms for the ICPC2 symptoms and diseases have been added to the ICPC2 ontology to integrate it with the consumer-oriented terminology.

5 Concluding Remarks and Future Works

This paper proposed a thesis work aiming at the creation of a consumer-oriented lexical-ontological resource that would help fill in the medical linguistic gap between specialized and "lay" terminology, and which could be used in consumer-oriented halthcare systems to help consumers in accessing to and managing of their healthcare data. In particular, preliminary results have been presented for the task of consumer-oriented terminology acquisition, on the basis of statistical and mapping analyses, which helped to find overlaps between extracted "lay" terms and specialized medical concepts in the ICPC2 medical terminology. First results are encouraging because many consumer-oriented terms were acquired, and a low overlap with ICPC2 medical concepts and a high number of synonyms were found. The formalization in terms of OWL axioms of the ICP2 and ICD10 coding systems, and the existing clinical mappig between them, were provided, and results were very positive allowing the reduction of the efforts for upgrading mappings in view of the next publication of the two encoded systems, ICD11 and ICPC3; and to reuse Mapping Consistency, Debugging, and Entailment.

This thesis work will potentially contribute to the state of the art in several research areas, including Medical Terminology, Healthcare Informatics, Knowledge Acquisition and Representation, and it can have the following potential aspects: 1) the Cross-Domain Interdisciplinarity; 2) the Integration of specialized and consumer-oriented medical knowledge, which helps to fill the medical communication gap; and 3) new methodologies for integration tasks and for knowledge services, which this framework will offer in the application for example to a PHR, to improve its management and accessibility.

[2] These ontologies can be consulted at: `https://dkm.fbk.eu/index.php/Resources`

To improve the results of the knowledge acquisition process and to extract more variegated consumer-oriented terminology, a written corpus, which include forum postings of an Italian medical website for asking questions to on-line doctors[3] has been analyzing. This will allow extending our sample and cover a wider range of ages, people with different background and consequently different levels of health literacy.

References

1. Bartolini, R., Lenci, A., Marchi, S., Montemagni, S., Pirrelli, V.: Text-2-knowledge: Acquisizione semi-automatica di ontologie per l'indicizzazione semantica di documenti. Technical Report for the PEKITA Project, ILC. Pisa p.23 (2005)
2. Bechhofer, S., Van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F., Stein, A.L.: OWL Web Ontology Language Reference, W3C Recommendation (2004)
3. Cardillo, E., Eccher, C., Tamilin, A., Serafini, L.: Logical Analysis of Mappings between Medical Classification Systems. In: Dochev, D., Pistore, M., Traverso, P. (eds.) AIMSA 2008. LNCS (LNAI), vol. 5253, pp. 311–321. Springer, Heidelberg (2008)
4. Cardillo, E., Serafini, L., Tamilin, A.: A Hybrid Methodology for Consumer-oriented Healthcare Knowledge Acquisition. In: proceedings of the KR4HC 2009 Workshop, Verona, July 19 (2009)
5. Ceusters, W., Smith, B., De Moor, G.: Ontology-Based Integration of Medical Coding Systems and Electronic Patient Records. In: MIE 2005 (2005)
6. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Heidelberg (2007)
7. Keselman, A., Logan, R., Smith, C.A., Leroy, G., Zeng, Q.: Developing Informatics Tools and Strategies for Consumer-centered Health Communication. Journal of Am. Med. Inf. Assoc. 14(4), 473–483 (2008)
8. Mork, P., Bernstein, P.: Adapting a generic Match Algorithm to Align Ontologies of Human Anatomy. In: Proceedings of ICDE (2004)
9. Noy, N.F., Rubin, D.L.: Translating the Foundational Model of Anatomy into OWL, in Web Semantics: Science, Services and Agents on the World Wide Web. Elsevier Science 6(2), 133–136 (2008)
10. Noy, N.F., Musen, N., Shah, N., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Montegut, M., Rubin, D., Youn, C.: BioPortal: A Web Repository for Biomedical Ontologies and Data Resources. In: The International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany (2008)
11. Rector, A.: Clinical Terminology: Why is it so hard? Methods of Information in Medicine 38(4), 239–252 (1999)
12. Rosembloom, T.S., Miller, R.A., Johnson, K.B., Elkin, P.L., Brown, H.S.: Interface Terminologies: Facilitating Direct Entry of Clinical Data into Electronic Health Record Systems. Journal of Am. Med. Inf. Assoc. 13(3), 277–287 (2006)
13. Zhang, S., Bodenreiden, O.: Experience in aligning anatomical ontologies. International Journal on Semantic Web and Information Systems 3(2), 1–26 (2007)
14. Zeng, Q., Goryachev, S., Keselman, A., Rosendale, D.: Making Text in Electronic Health Records Comprehensible to Consumers: A Prototype Translator. In: The 31st American Medical Informatics Association's Annual Symposium, AMIA 2007, pp. 846–850 (2007)

[3] http://medicitalia.it