A Lexical-Ontological Resource for Consumer Healthcare[*]

Elena Cardillo, Luciano Serafini, and Andrei Tamilin

FBK-IRST, Via Sommarive 18, 38050 Povo (Trento), Italy
{cardillo,serafini,tamilin}@fbk.eu

Abstract. In Consumer Healthcare Informatics it is still difficult for laypeople to find, 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 that will be integrated with other medical domain ontologies and terminologies into a medical ontology repository. This will support consumer-oriented healthcare systems, such as Personal Health Records, by providing many knowledge services to help users in accessing and managing their healthcare data.

Keywords: Medical Vocabulary Acquisition, Consumer-oriented Terminologies, Healthcare Ontologies.

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 patients and healthcare consumers in general. The medical communication gap became more evident when consumers started to play an active role in healthcare information access. In fact, they have become more responsible for their personal healthcare data, 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. To help consumers fill this gap, the challenge is to sort out the different ways consumers communicate within distinct discourse groups and map the common,

[*] This work is supported by the TreC Project, funded by the Province of Trento.

P. Kostkova (Ed.): eHealth 2009, LNICST 27, pp. 131–138, 2010.

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, 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, facilitating the process of codification of symptoms, diagnoses and diseases, there is little work based on the use of consumer-oriented medical terminology, moreover most of existing studies focused only on English.

Given this scenario, we want to propose a methodology for the creation of a lexical-ontological resource oriented to healthcare consumers in the Italian context, and its integration with a coherent semantic medical resource, useful both for professionals and for consumers. Such a resource can be used in healthcare systems, like Personal Health Records (PHRs), as to help consumers during the process of querying and accessing healthcare information, so as to bridge the communication gap. In the present work we focus in particular on the use of a hybrid methodology for the acquisition of consumer-oriented "lay" terminology for expressing medical concepts such as symptoms, diseases, anatomical concepts, for the consequent creation of a Consumer Medical Vocabulary for Italian, which can be used to translate technical language with lay terminology and vice-versa.

2 Medical Terminologies and Ontologies

Over the last two decades research on Medical Terminologies has become a popular topic and the standardization efforts have established a number of terminologies and classification systems, such as SNOMED International[1] or ICD-10 (International Classification of Diseases)[2], as well as conversion mappings between them to help medical professionals in managing and codifying their patients health care data. They concern with *"the meaning, expression, and use of concepts in statements in the medical records or other clinical information systems"* [6]. Having all these medical terminologies interoperability has become a significant problem. Content, structure, completeness, detail, cross-mapping, taxonomy, and definitions vary between existing vocabularies. During the last few years, thanks to the Semantic Web perspective, a set of new methodologies and tools were generated for improving healthcare systems, in particular translating medical terminologies into more formal representations using ontology methodologies and languages (e.g., the formalization of SNOMED CT [7]).

Much effort has also been spent for the creation of new Biomedical Ontologies, such as the Foundational Model Anatomy (FMA) [5], and GALEN into OWL. Ontologies in the medical domain provide an opportunity to leverage the capabilities of OWL semantics and tools to build formal, sound and consistent medical terminologies, and to provide a standard web accessible medium

[1] `http://www.ihtsdo.org/snomed-ct/`

[2] `http://www.who.int/classifications/icd/en/`

for interoperability, access and reuse. Given the presence of all these medical ontologies, two other important issues have to be taken into account: Ontology Mapping, to show how concepts of one ontology are semantically related to concepts of another ontology; and Ontology Integration, which allows access to multiple heterogeneous ontologies[3].

Despite these advantages, the vocabulary problem continues to plague health professionals and their information systems, and especially laypeople, who are the most damaged by the increased communication gap. To respond this consumer needs, during the last few years, many researchers have labored over the creation of lexical resources that reflect the way healthcare consumers express and think about health topics. One of the largest initiatives in this direction is the Consumer Health Vocabulary Initiative[4], by Q. Zeng and colleagues at Harvard Medical School, 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 in Keselman *et al.* [3].

It is important to stress that there are only few examples of the application as far as these initiatives are concerned. For example, in Kim *et al.* [4] and Zeng *et al.* [9] we find an attempt to face syntactic and semantic issues in the effort to improve PHRs readability, using the CHV to map their content. On the other hand, Rosembloom *et al.* [8] 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 Approach

The global approach followed for this research activity is divided in two macro phases. The first one includes the creation of a Consumer Medical Vocabulary (CMV) for Italian, for collecting common medical expressions and terms used by Italian speakers. The second one focuses on the formal representation of some relevant medical terminologies, which will be integrated with the CMV, and the development of a Medical Ontology Repository (MORe) in which all these ontologies and terminologies will be integrated. This activity can be further 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.

[3] http://www.obofoundry.org

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

- Creation of the Italian Consumer Medical Vocabulary (CMV). Selection of all the lay terms extracted that have been identified as good representatives for medical concepts (considered as synonyms after clinical review performed by physicians), 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 integrated with the CMV and some existing medical ontologies, relevant for our aims, 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.

In this work we focus on the task of acquisition of consumer-oriented knowledge about a specific subset of healthcare domain, and the creation of the consumer-oriented medical vocabulary for Italian. Concerning the task of Formalization of medical terminologies we refer the reader, for the details of the applied methodology and preliminary results, to Cardillo *et al.* [2]. There we present the formalization of two medical classification systems, ICPC2 and ICD10, into OWL ontologies. In the same work we describe also the construction of a well-founded and medically sound mapping model between the two ontologies by means of the formalization in terms of OWL axioms of the existing clinical mappings, and validation of its coherence using Semantic Web techniques.

3.1 Knowledge Acquisition Task

We used a hybrid methodology for the identification of “lay” terms, words, and expressions used by Italian speakers to indicate Symptoms, Diseases, and Anatomical Concepts. Three different target groups were considered: First Aid patients subjected to a Triage Process; a community of Researchers and PhD students with a good level of healthcare literacy, and finally a group of elderly people with a modest background and low level of healthcare literacy. This methodology consisted of the following steps:

1. Application of three different Elicitation Techniques to the mentioned groups;
2. Automatic Term Extraction and analysis of acquired knowledge by means of a Text Processing tool;
3. Clinical review of extracted terms and manual mapping to a standard medical terminology (ICPC2), performed by physicians;
4. Evaluation of results in order to find candidate terms to be included in the Consumer-oriented Medical Vocabulary.

Wiki-based Acquisition. The first method is based on the use of a Semantic Media Wiki system, an easy to use collaborative tool, allowing users to create and link, in a structured and collaborative manner, wiki pages on a certain domain of knowledge. Using our online *eHealthWiki* system[5], users created wiki pages for describing symptoms and diseases, using “lay” terminology, specifying in particular the corresponding anatomical categorization, the definition and possible synonyms. The system has been evaluated over a sample of 32 people: researchers, PhD students and administrative staff of our research institute (18 females, 14 males, between 25 and 56 years old). In one month, we collected 225 wiki pages, 106 for symptoms and 119 for diseases, and a total of 139 synonyms for the inserted terms. Users were reluctant to the collaborative functionality of this system, which allows modifying concepts added by others.

Nurse-assisted Acquisition. The second method involved nurses of a First Aid Unit[6] as a figure of mediation for the acquisition of terminology about patient symptoms and complaints, helping them to express their problems using the classical subjective examination performed during the Triage Process, which aims to prioritize patients based on the severity of their condition. This method involved 10 nurses, around 60 patients per day and a total of 2.000 Triage Records registered in one month. During this period nurses acquired the principal problems (symptoms and complaints) expressed by their patients using “lay” terminology and inserted them in the Triage Record together with the corresponding medical concepts used for codifying patient data.

Focus-Group Acquisition. The method consisted in merging the following elicitation techniques: Focus Group, Concepts Sorting, and Board Games, in order to allow interaction and sharing situations to improve the process of acquisition. The target was a community of 32 elderly people in a Seniors Club, between 65 and 83 year old. We used group activities to acquire, even in this case, lay terms and expressions for symptoms, diseases and anatomy. About 160 medical terms were collected. Then all the terms were analyzed together with other groups, creating discussions, exchanging opinions on terms definitions, synonyms, and recording preferences and shared knowledge. At the end, all participants gave preferences for choosing the right body system categorization (digestive, neurological, musculoskeletal, lymphatic, endocrine, etc.) of each of the written concept.

3.2 Term Extraction and Mapping Analysis

The three sets of collected data were further processed and analyzed, to detect candidate consumer-oriented terms, with Text-2-Knowledge tool (T2K) [1]. This tool allowed us to automatically extract terminology from the data sets, to perform many text processing techniques, and to calculate statistics on the extracted data such as term frequency. In spite of the advantages of the automatic extraction process, allowing for extraction of many compound terms, such a procedure has demonstrated that a large amount of terms, certainly representative of consumer medical terminology, were not automatically extracted, since, due

[5] `http://ehealthwiki.fbk.eu`

[6] `http://www.apss.tn.it/Public/ddw.aspx?n=26808`

to the quantitative limits of the corpus dimensions, their occurrence was inferior with respect to the predefined threshold value. Consequently, we performed an additional manual extraction to take into account such rare terms, usually mentioned by a single participant.

Extracted terms were reviewed by two physicians to find incongruences in categorization and synonymy. For instance, “Giramento di Testa” (*Dizziness*) was categorized as Cardiovascular problem instead of Neurological. Physicians have been also asked to map a term/medical concept pair by using a professional health classification system, the above mentioned ICPC2. It addresses fundamental parts of healthcare process: it is used in particular by general practitioners for encoding symptoms and diagnosis. We identified five different types of relations between consumer terms and ICPC2 medical concepts:

- Exact mapping between the pairs; this occurs when the term used by a lay person can be found in ICPC2 rubrics and both terms correspond to the same concept. For example, the lay term “Febbre” (*Fever*) would map to a ICPC2 “Febbre” term, and both will be rooted to the same concept.
- Related mapping; it involves lay synonyms and occurs when the lay term does not exist in the professional vocabulary, but corresponds to a professional term that denotes the same (or closely related) concept. E.g., lay term “Sangue dal Naso” (*Nosebleed*) corresponds to “Epistassi” (*Epistaxis*).
- Hyponymy relation; this occurs when a lay term can be considered as term of inclusion of a ICPC2 concept. E.g., lay term “Abbassamento della Voce” (*Absence of Voice*) is included in the more general ICPC2 concept “Sintomo o disturbo della voce” (*Voice Symptom/Complaint*).
- Hyperonymy relation; in this case the lay term is more general than one or more ICPC2 concepts, so it can be considered as its/their hyperonym. E.g., the term “Bronchite” (*Bronchitis*) is broader than “Bronchite Acuta/ Bronchiolite” (*Acute Bronchitis/ Bronchiolitis*) e “Bronchite Cronica” (*Chronic Bronchitis*) ICPC2 concepts.
- Not mapped; those lay terms that cannot be mapped to the professional vocabulary. These can be legitimate health terms, the omission of which reflects real gaps in existing professional vocabularies; or they can represent unique concepts reflecting lay models of health and disease. E.g., the lay term “Mal di mare” (*Seasickness*).

4 First Results Evaluation

We were able to acquire a variegated consumer-oriented terminology and to perform an interesting terminological and conceptual analysis. By means of the term extraction process, from 225 Wiki pages, we were able to extract a total of 962 medical terms. We found a total of 173 Exact Mappings, 80 Related Mappings, 94 Hyperonyms, 51 Hypomyms and, finally, 186 Not Mapped ICPC2 concepts. Most of the exact mappings with ICPC2 are related to anatomical concepts, and many synonyms were found for symptoms. Concerning the Nurse-assisted data set, from 2.000 Triage records we extracted a total of 2389 terms, but about half of these terms were considered irrelevant for our evaluation, so mapping was provided only for 1108 terms. Here we can highlight 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 transcription of the Focus Group/Game activity (third data set). Here all the symptoms extracted (79 terms) had a corresponding medical concept in ICPC2 terminology (35 Exact Mappings and 44 Related Mappings).

Table 1 below compares the three data sets together and shows that the most profitable method for acquiring consumer-oriented medical terminology was the one assisted by Nurses. Also Wiki-based method, even if not exploited for the collaborative characteristic, has demonstrated good qualitative and quantitative results. Results concerning mapping to ICPC2 can be considered, because 2/3 of the terms extracted are covered by ICPC2 terminology. 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 we extracted a very variegated range of medical terms, many compound terms and expressions, which can be representative for the corresponding technical ones in standard terminology, and which can be used as candidate for the construction of our Consumer-oriented Medical Vocabulary for Italian.

Table 1. Mapping Results

Sources	Total Terms	Mapped	Not Mapped
Wiki-based	962	398	186
Nurse-assisted	2389	726	382
Focus-Group	321	231	12
Total	3662	1355	580

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

In this paper we proposed the creation of a lexical-ontological resource for healthcare consumers that would help to fill in the linguistic communication gap between specialized and "lay" terminology. Such a resource can be used in consumer-oriented healthcare systems in order to help users in accessing to and managing of their healthcare data. We have presented preliminary results for the task of consumer-oriented terminology acquisition, on the basis of statistical and mapping analysis, which helped us to find overlaps between extracted “lay” terms and specialized medical concepts in the ICPC2 terminology. Our methodology showed encouraging results, because it allowed us to acquire many consumer-oriented terms; a low overlap with ICPC2 and a high number of related mappings (mainly synonyms) to the referent medical terminology. To improve the results of the acquisition task and to extract more variegated consumer-oriented terminology, not related to the regional context, we are analyzing written corpora, which include forum postings of an Italian medical website for asking questions to on-line doctors [7]. This will allow extending our sample and cover a wider range of ages, people with different background and consequently different levels of health literacy. This task will be very interesting for comparing results with that came out from the previous elicitation methods, both in quantitative and qualitative terms.

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, 23 (2005)
2. Cardillo, E., Eccher, C., Tamilin, A., Serafini, L.: Logical Analysis of Mappings between Medical Classification Systems. In: Proc. of the 13th Int. Conference on Artificial Intelligence: Methodology, Systems, and Applications, pp. 311–321 (2008)
3. Keselman, A., Logan, R., Smith, C.A., Leroy, G., Zeng, Q.: Developing Informatics Tools and Strategies for Consumer-centered Health Communication. Journal of American Medical Informatics Association 14(4), 473–483 (2008)
4. Kim, H., Zeng, Q., Goryachev, S., Keselman, A., Slaughter, L., Smith, C.A.: Text Characteristics of Clinical Reports and Their Implications for the Readability of Personal Health Records. In: Proc. of the 12th World Congress on Health (Medical) Informatics, MEDINFO 2007, pp. 1117–1121 (2007)
5. 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 6(2), 133–136 (2008)
6. Rector, A.: Clinical Terminology: Why is it so hard? Methods of Information in Medicine 38(4), 239–252 (1999)
7. Rector, A., Brandt, S.: Why do it the hard way? The case for an expressive description logic for SNOMED. Journal of American Medical Informatics Association 15(6), 744–751 (2008)
8. 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 American Medical Informatics Association 13(3), 277–287 (2006)
9. Zeng, Q., Goryachev, S., Keselman, A., Rosendale, D.: Making Text in Electronic Health Records Comprehensible to Consumers: A Prototype Translator. In: Proc. of the 31st American Medical Informatics Association’s Annual Symposium, AMIA 2007, pp. 846–850 (2007)

[7] `http://medicitalia.it`