international journal of medical informatics 76 (2007) 208–215

journal homepage: www.intl.elsevierhealth.com/journals/ijmi

Building an ontology of pulmonary diseases with natural language processing tools using textual corpora

Audrey Baneyx[a,*], *Jean Charlet*[a,b], *Marie-Christine Jaulent*[a]

[a] *INSERM U729, Laboratoire SPIM, Faculté de Médecine Broussais, Hôtel-Dieu, 15 rue de l'Ecole de médecine, Paris F-75006, France*
[b] *STIM DSI/AP-HP, Paris F-75014, France*

ARTICLE INFO

Article history:
Received 3 January 2006
Received in revised form
20 March 2006
Accepted 2 May 2006

Keywords:
Ontology
Terminology extraction
Knowledge engineering
Thesaurus
Natural language processing
Pulmonary diseases

ABSTRACT

Pathologies and acts are classified in thesauri to help physicians to code their activity. In practice, the use of thesauri is not sufficient to reduce variability in coding and thesauri are not suitable for computer processing. We think the automation of the coding task requires a conceptual modeling of medical items: an ontology. Our task is to help lung specialists code acts and diagnoses with software that represents medical knowledge of this concerned specialty by an ontology. The objective of the reported work was to build an ontology of pulmonary diseases dedicated to the coding process. To carry out this objective, we develop a precise methodological process for the knowledge engineer in order to build various types of medical ontologies. This process is based on the need to express precisely in natural language the meaning of each concept using differential semantics principles. A differential ontology is a hierarchy of concepts and relationships organized according to their similarities and differences. Our main research hypothesis is to apply natural language processing tools to corpora to develop the resources needed to build the ontology. We consider two corpora, one composed of patient discharge summaries and the other being a teaching book. We propose to combine two approaches to enrich the ontology building: (i) a method which consists of building terminological resources through distributional analysis and (ii) a method based on the observation of corpus sequences in order to reveal semantic relationships. Our ontology currently includes 1550 concepts and the software implementing the coding process is still under development. Results show that the proposed approach is operational and indicates that the combination of these methods and the comparison of the resulting terminological structures give interesting clues to a knowledge engineer for the building of an ontology.

1. Introduction

For about 10 years, French public hospitals have had to communicate information about their medical activities. For each patient, information is gathered as patient discharge summary, using the international classification of diseases CIM-10[1] for diagnoses codification and CCAM[2] for acts codification. The French PMSI[3] coding process is usually done manually by

* *Corresponding author*.
Tel.: +33 1 53 10 92 12 (O); +33 6 62 77 59 85 (R)
E-mail addresses: Audrey.Baneyx@spim.jussieu.fr (A. Baneyx), Jean.Charlet@spim.jussieu.fr (J. Charlet), Marie-Christine.Jaulent@spim.jussieu.fr (M.-C. Jaulent).

1386-5056/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2006.05.031

[1] The French version of the international classification of diseases : http://www.med.univ-rennes1.fr/noment/cim10/.

[2] Common classification of medical acts : http://www.codage.ext.cnamts.fr/codif/ccam/index_presentation.php.

[3] Program of medicalization of information systems: http://www.ch-charcot56.fr/dossiers/pmsi/pmsi.htm.

physicians using medical specialty thesauri based on common terminologies. It has become obvious that the use of such thesauri is not sufficient to reduce variability in coding. Indeed, wording of thesauri is ambiguous (for instance several pathologies apply to a unique code) and non-exhaustive; the chosen classification method is difficult to use; lastly, maintaining either consistency or coherence of thesauri is impossible [1]. Moreover, the interpretation of wording of medical terminologies depends on the human reader, and as such is not adapted to computer processing. Some work has been done on automatic coding tools to reduce mistakes and variability in coding [2]. However, it has been argued in the literature that the automation of the coding task requires a conceptual organization of medical items whose meaning would be written inside the model's structure itself [3] within an "ontology" [4]. The word "ontology" introduced in philosophy has been reused since the beginning of the 1990's mainly in artificial intelligence, knowledge engineering and knowledge representation. Today, its scope is growing and it is becoming a common item in the field of information-system modeling [5]. An ontology is a formal system whose purpose is to represent knowledge in a specific domain by means of basic elements called concepts, which are defined and organized in relation to the one another [6,7]. An ontology ensures that the coherence of the axioms and the integrity of the system as well as the extensibility of the structure is maintained. The main difficulty is to identify and classify the items of a given domain. Since classification criteria depend on purposes and are not universal, we do not seek to build a universal ontology, but merely a specific ontology of pulmonary diseases [8,9]. We assume that the development of ontological resources will allow us to reach high-performance, reliable and progressive coding tools.

This paper is organized as follows: Section 2 details our objectives. Section 3 introduces the material and tools used for the work. Section 4 describes the different steps of the methodology. Section 5 presents the results and, in Section 6, we conclude this paper by discussing perspectives expected from this work.

2. Objectives

This work is part of the PERTOMed[4] research project whose objective is to develop an internet platform offering a range of methods and tools to produce and use terminological and ontological resources in the medical field. Specific medical ontologies are created in close partnership with user groups, who will participate in the evaluation processes of these resources in their real use environment. Our task in the project is to help lung specialists code acts and diagnoses with software that represents medical knowledge by an ontology of the concerned specialty. We are working with the French Pneumology Society.[5] We notice there is no ontology covering the pulmonary diseases field designs for French coding process. The objective of the present work is the building of ontology for this purpose. Many approaches have been reported to build ontologies (for a complete survey, the reader can refer to the OntoWeb Technical RoadMap[6]), but few fully detail the conceptualization steps, in which concepts and relationships are captured and organized. The main constraint of our work is to have this ontology built by a knowledge engineer rather than directly by a physician. The main difficulty for a knowledge engineer is then to identify and classify the concepts of a given domain. We apply a text-driven knowledge approach and consider textual reports as the main source of information. Natural language processing (NLP) tools are applied to analyse corpora. We believe that a method for the building of a well-formed ontology has to assign a clear meaning to concepts. One of our work hypothesis is that the most natural way to precise each concept is to explain its meaning in natural language [3,10]. In fact, the words used to denote concepts are still liable to multiple interpretations. This results in possible misunderstandings and consequently bad modeling and use of the ontology. As a solution, the method we use as a guide is based on differential semantics principles [9]. The ontology is structured by nodes forming a tree. The main difference between regular and differential ontologies lies in the importance they attach to locate precisely a concept in the ontological tree and the means given to do it. In the differential case, the meaning of a node is given by the gathering of all similarities and differences, expressed in natural language, between the node target and the root. In other words, the meaning of a concept is determined by its position in the ontological hierarchy based on terminological structure. In this paper we describe an experiment in building the ontology of pulmonary diseases. We develop a precise methodological process for the knowledge engineer in order to build various types of medical ontologies according to the differential semantics principles inherited from the Differential Semantics Theory by Rastier et al. [11]. This process requires medical experts only at particular validation times. The originality of our approach consists in applying NLP tools to corpora to develop the resources needed to build the ontology. Our main research hypothesis concerns the joint use of two methods to enrich the ontology building: (i) a method which consists of building terminological resources by distributional analysis [12,13], and (ii) a method based on semantic relationship recognition by the observation of corpus sequences [14,15].

3. Material

In order to cover as exhaustively as possible the whole area of pulmonary diseases, we have gathered 1038 patient discharge summaries (corpus named [PDS]) from six hospitals of the Assistance Publique-Hôpitaux in Paris, France. In a previous work, it has been shown that 350,000 words seems to be a good indicator [16]. In the pulmonary diseases case, this first corpus [PDS] has about 417,000 words, which seems to be a good basis for the experiment. We added a teaching book to that first corpus (corpus named [BOOK]), which enabled us

[4] http://pertomed.spim.jussieu.fr.

[5] http://www.splf.org.

[6] http://babage.dia.fi.upm.es/ontoweb/wp1/OntoRoadMap/index.html.