Building Medical Ontologies Based on Terminology Extraction from Texts: an Experimentation in Pneumology

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

[a]*INSERM U729, Paris, F-75006, France;*
[b]*STIM DSI/AP-HP, Paris, F-75014, France;*

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 do not fit computer processing. We think the automation of the coding task requires a conceptual modelling of medical items: an ontology. Our objective is to help pneumologists code acts and diagnoses with a software that represents medical knowledge by an ontology of the concerned specialty. The main research hypothesis is to apply natural language processing tools to corpora to develop the resources needed to build the ontology. In this paper, our objective is twofold: we have to build the ontology of pneumology and we want to develop a methodology for the knowledge engineer to build various types of medical ontologies based on terminology extraction from texts.

Keywords:
Ontology, Terminology Extraction, Knowledge Engineering, Differential Semantics, Natural Language Processing, Pneumology.

1. Introduction

In many medical specialities, pathologies and acts are classified in thesauri to help physicians to code their activity. 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 as 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. Moreover, the interpretation of wording of medical terminologies depends on the human reader, and as such are not adapted to computer processing. We think that the automation of the coding task requires a conceptual modelling (an ontology non-contextual and unambiguous) of medical items whose meaning would be written inside the model's structure itself. Such modelling is called "ontology" [1]. The word "ontology" used in philosophy has been reused since the beginning of the 1990's mainly in artificial intelligence, knowledge engineering and knowledge representation. Nowadays, its scope is growing and it becomes a common item in the field of information-system modelling [2]. An ontology is a formal system whose objective is to represent a specific domain knowledge by means of basic elements, the concepts, that are defined and organised each one in relation to the others [3]. Ontology insures to maintain axioms' coherence and system's integrity as well as extensibility of the structure without modifying it. 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 pneumology. This work is part of the PERTOMed[1] 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 pneumologists code acts and diagnoses with a software that represents medical knowledge by an ontology of the concerned specialty. The main research hypothesis of PERTOMed is to apply natural language processing (NLP) tools to corpora to develop the resources needed to build the ontology. Thus, an important task is to create textual corpora necessary for NLP tools.

In this paper, the objective is twofold. On the one hand, we have to build the ontology of pneumology and, on the other hand, we want to contrive a methodology for the knowledge engineer to build various types of medical ontologies based on the differential semantics theory [4].

Section 2 presents the material and tools used for the experimentation. Section 3 details the different steps of the methodology. Section 4 presents the results we obtained and last, in section 5, we conclude this paper by discussing perspectives expected from this work.

2. Material

Presentation of the two corpora

In order to cover the whole area of pneumology, we have gathered patient discharge summaries (PDS) in six hospitals of the Assistance Publique-Hôpitaux de Paris. We ended up with a total of 1 038 PDS. In a previous work about a different domain, it has been showed that 350 000 words seems to be a good indicator [5]. In the pneumology case, the first corpus [PDS] has about 417 000 words which is a good base for the experimentation. We added a teaching book to that first corpus, which permitted to refine and control the hierarchical structure of the ontology during the development. That second corpus [BOOK] has about 823 000 words.

Available tools

We use SYNTEX-UPERY as an NLP tool. SYNTEX is a syntax analyser module allowing us to get a network of syntactic dependencies between terms - or terminological network - whose elements are the candidate terms that will be used for the building of the ontology. UPERY is a distributional analyser module computing distributional proximities between candidate terms on the basis of shared syntaxical contexts. It exploits all the network data to cluster needed terms. We obtain a network of candidate terms, their contextual associations and their links to the corpus. The differential ontology editor DOE[2] allows us to build our ontology according to the differential semantics theory proposed in [4].

3. Method

The Bachimont's methodology used to design differential ontologies allows us to describe variations of the words' meaning in context. It stresses the importance of the textual corpus [6], since that corpus is the best source to characterize notions useful for the ontological modelling and the semantic content that is associated with them. We distinguish four successive steps: 1) the constitution of the corpus of knowledge and its analysis by NLP

[1] http://www.spim.jussieu.fr/Pertomed
[2] http://opales.ina.fr/public/

tools, 2) the semantic normalization of the set of terms through the application of differential principles, 3) the ontological engagement to formalize defined concepts, 4) the finalization of the process in a language, based on description logics, understandable by the computer [4]. Our experimentation in the pneumology specialty brings both precision and adapts the first two steps of the methodology to the knowledge engineer which is a contribution of our second objective.

Processing of basic resources

The two [PDS] and [BOOK] corpora resources are first processed in order to obtain an anonymous [PDS] corpus, and a didactical [BOOK] corpus, both in XML format. They are processed by SYNTEX-UPERY. The result of the analysis of candidate terms in the [PDS] corpus allows us to build the basic elements – i.e. primitives – of the ontology. A candidate term is a noun phrase (NP) composed by a head and an expansion. For instance, in the NP *Opacity in the left lung,* the term *Opacity* is the head and *in the left lung* is its expansion. The results of the analysis of the [BOOK] corpus are processed themselves to identify hyperonymy links between candidate terms [6], using lexico-syntactic patterns. Those links help us to structure the hierarchy of primitive concepts.

Building of the ontology, stage 1: choice of the candidate terms

Candidate terms representative of pneumology are chosen within the results provided by SYNTEX-UPERY from the [PDS] corpus in two steps:

1) We scan all the results provided by the syntaxical analysis and choose to first study the NP that appear in the corpus more than 12 times (2% of the corpus). From then, we spot the major conceptual axes that are typical of the corpus and the medical field. To each candidate term, we associate a validity criterion that matches one of those axes, on a 1 to 6 scale: 1 (non pertinent terms, axis: other), 2 (reserved for terms which are modellised in the ontology), 3 (axis: symptoms), 4 (axis: pathologies), 5 (axis: treatments) and 6 (axis: examinations). For instance, we associate the validity criterion 6 with all candidate terms on that axis – e.g. *echography*, *doppler*, *ultrasonography*, and so on… At the beginning of the method, all the candidate terms have a validity criterion equal to 1 and at the end they are all classified on the axis 2 (which means they have been completely defined in the ontology). Validity criteria 3, 4, 5 and 6 are used temporarily during the building phase (stage 2). This gathering allows the beginning of a first stage of work on the connection by context and the selection by validity criteria leaves 35% of candidate terms on which we can work out the heart of our ontology.
2) The distributional analysis connects terms sharing the same contexts (descendants in head and descendants in expansion). It also connects the contexts according to the terms they share (neighbours in head and neighbours in expansion). On the example (Cf. Table 1), *effusion* is the head of the NP *effusion of pleura* and *of pleura* is its expansion. Descendants in head yield information on what could be child-concepts or defined concepts. Descendants in expansion provide information about the concept's position in the hierarchy. Neighbours in head and in expansion allow us to constitute the groupings of candidate terms semantically close to the one under study, *effusion of pleura*. Groupings are a great help for the development of the hierarchical structure of the ontology, for both horizontal and vertical axes. The example below shows a first possible connection: we can link group A *{effusion, lesion, infection, uncompensation}* with *{symptoms}*. The candidate terms of group A share the same semantical context. Our first hypothesis will be to consider the candidate terms of group A as sibling-concepts and *symptoms* as the parent-concept.

Table 1 – Sample results of the contextual connections for the NP "Effusion of pleura"

Descendants in head	Descendants in expansion	Neighbours in head	Neighbours in expansion
Effusion of the right part of pleura	Tapping of pleural effusion	Lesions	Liquid
Effusion of liquid	Recurrence of pleural effusion	Infections	Infiltration
Effusion of the pleural cavity	Link between dyspnoea and pleural effusion	Symptoms	Uncompensation

Building of the ontology, stage 2: application of differential principles

In order to work the hierarchy out, the candidate terms chosen in stage 1 are organised by refining the differential principles that define them. We have to express in natural language the similarities and differences of each concept with respect to its parent-concept and its sibling-concepts. For instance, the concept of *Ultrasonography* and the concept of *IsotopicExamination* are sibling-concepts whose parent-concept is *RayImaging*. The principle of similarities with the parent-concept is the projection or the injection of an artificial substance in order to take measures. The principle of similarities between sibling-concepts is related to the injection medium. The differencial principle between sibling-concepts is related to the kind of artificial medium used: an isotope in the case of an isotopic examination and ultrasounds in the case of ultrasonography. The 4 axes (3, 4, 5, and 6) are refined that way. In addition, we use the results of the processing of the [BOOK] corpus to help us in applying the differential principles according to the method described in [4].

During the building process, the knowledge engineer must distinguish between primitive and defined concepts. A primitive concept is essential to the representation of the specialty. A defined concept is built from one or more primitive concepts and one or more relationships. The NP *left thoracic pain* is modellised by the following defined concept: [pain] (at the) [thorax] (located on) [left][3].

At the end of this stage, we have a semantic normalization of the set of terms of the specialty and we have represented the hierarchy of primitive concepts and relationships with DOE.

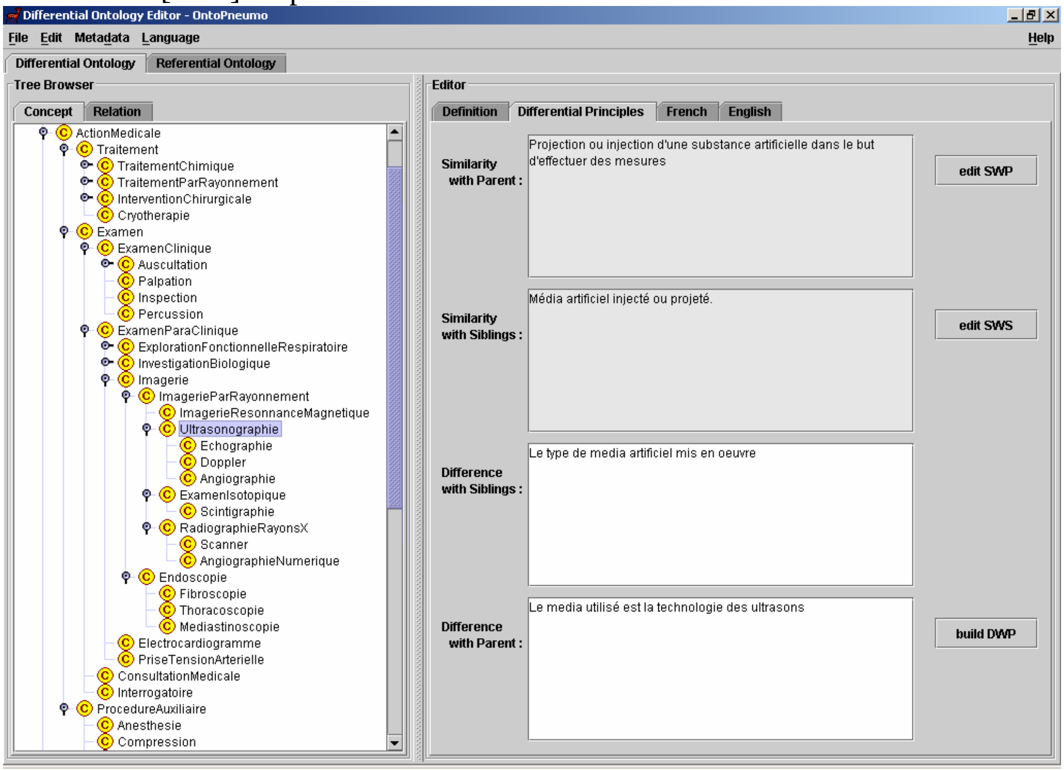
4. Results

Primitive concepts and their organisation

After using SYNTEX process, the [PDS] corpus gives 36 881 NP and the [BOOK] corpus gives 17 666 NP. According to the results of the syntaxical and distributional analyses, the NP *Chemotherapy of <noun>* has the most neighbours in head, 28, and has 90 appearances in the corpus. The NP *<noun> of chemotherapy* has the most neighbours in expansion, 52, and has the highest appearance rate, 454. We can check the pertinence of the connections by group of candidate terms whose appearance contexts are semantically close. For instance, *course of chemotherapy* has *{Hospitalization, Examination, Navelbine, Cisplatine, Taxotere, Carboplatine}* as its neighbours in head and has *{Treatment, Check-up, Antibiotherapy, Injection, Radiotherapy}* as its neighbours in expansion. These results are examined and then transformed into ontological form, using DOE. We build the following hierarchy: *MedicalAction/Treatment/DrugTreatment/Chemotherapy*. Chemotherapy being considered as a drug treatment, we find the following medicinal principles classified under *Medicine/…/MedicinalPrinciple/Navelbine, Cisplatine, Doxorubicine, Taxotere,*

[3] The square brackets indicate a primitive concept and the brackets indicate a relationship. In our methodology, relationships are processed in the same way as concepts.

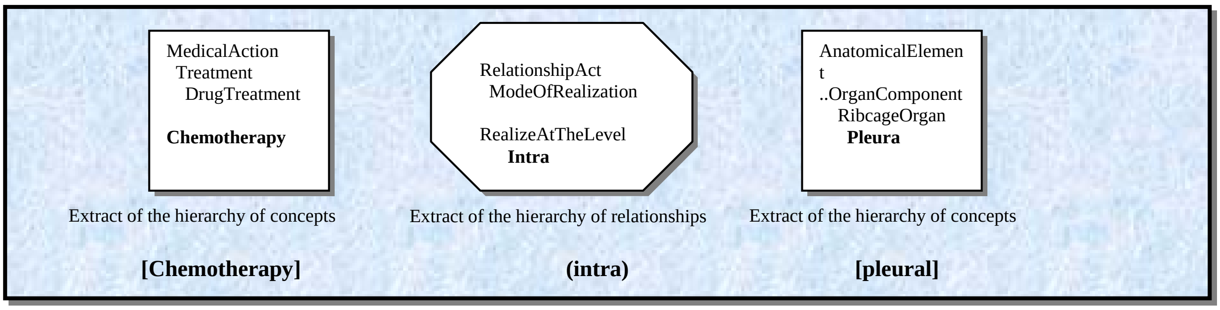
Carboplatine. The candidate terms *Antibiotherapy* and *Radiotherapy* are also located under *Treatment*. This method for regrouping yields an ontology and makes the task much easier for a knowledge engineer who would not be a specialist in the modellised medical specialty. Our ontology includes today 750 primitive concepts stemming from the first study of the candidate terms. Given that the building stages 1 and 2 are iterative, we will rapidely increase the hierarchy by examining the candidate terms which appear less than 12 times in the [PDS] corpus.



– An extract of the ontology of the pneumology viewed with DOE

Preliminary method of validation

The ontology is to be evaluated in terms of both quality and coverage. The conceptual hierarchy must be corrected and validated by pneumologists from the French Pneumology Society with whom we collaborate. The evaluation will also test the completeness of the ontology compared to the specialty thesaurus. To estimate that completeness, we will check the possibility of building a conceptual representation of medical knowledge by combining the primitive concepts and the relationships in the ontology. Today, we are able to build some phrases, for instance the phrase *Chemotherapy intra-pleural*, which is in the Pleura chapter in the thesaurus.



5. Discussion and conclusion

An important issue of this work is to show that our methodology allows a knowledge engineer, a non-specialist in the modellised field, to build ontology from texts, using NLP tools. As a new experimentation, we have tried to work out an organisation of the pneumology concepts, both coherent and stable, based on the practical application of the principles of differential semantics. The initial results of our research coupled with a recent work on surgical intensive care gives us reasons to think we are moving in the right direction [5]. We also forecast an extension of the ontology by analysing the terms specific to pneumology in the ICD-10[4] and CCAM[5] thesauri. We plan to complete this ontology by connecting it to the head (conceptual high level) of the Ménélas project ontology[6] [7]. We assume we will rapidly get an ontology of pneumology of approximately 1 200 concepts. The validation of the ontology can only be done through its use in concrete applications. It must be made available for the medical profession through an environment for helping the coding of acts and diagnoses and the representation of medical knowledge, as part of the terminological platform of the PERTOMed project.

6. Thanks and acknowledgments

We would like to particularly thank Pr Thomas Similowski for his help, Mr Didier Bourigault for making available SYNTEX-UPERY, Mrs Raphaël Troncy and Antoine Isaac who helped us in the use of DOE. We also wish to thank Pr Bruno Housset (Service de Pneumologie, Centre Hospitalier InterCommunal de Créteil, et Société de Pneumologie de Langue Française), Pr Bernard Maitre (Unité de Pneumologie, CHU Henri-Mondor, Créteil), Dr François-Xavier Blanc (Unité de Pneumologie, Service de Médecine Interne, CHU du Kremlin-Bicêtre, le Kremlin-Bicêtre), Dr Nicolas Roche (Service de Pneumologie, Hôtel-Dieu, Paris), Pr Christos Chouaid (Service de Pneumologie, Hôpital St-Antoine, Paris), Pr Jacques Cadranel (Service de Pneumologie, Hôpital Tenon, Paris), Pr Marc Humbert (Service de Pneumologie, Hôpital Antoine Béclère, Clamart), and Dr Alexandre Duguet (Unité de Réanimation du Service de Pneumologie, Groupe Hospitalier Pitié-Salpêtrière, Paris) for having taken time to collect the resources necessary for our work.

7. References

[1] Staab S and Studer R. *Handbook on Ontologies.* 1st ed. Berlin: Springer, Handbooks in Information Systems, 2003.

[2] Gomez-Pérez A, Fernandez-Lopez M, and Corcho O, eds. *Ontological Engineering.* 2nd ed. Madrid: Springer, Advanced Information and Knowledge Processing, 2004.

[3] Rector A. L. Thesauri and Formal Classifications: Terminologies for People and Machines. *Methods of Information in Medicine, 1998.* 37: 501-9.

[4] Bachimont B, Isaac A, and Troncy R. Semantic Commitment for Designing Ontologies: A Proposal. In Gomez-Pérez A and Richard Benjamins V, eds. *EKAW'2002.*

[5] Le Moigno S, Charlet J, Bourigault D, Degoulet P, and Jaulent M-C. Terminology extraction from text to build an ontology in surgical intensive care. *AMIA, 2002*. 9-13.

[6] Malaisé V, Zweigenbaum P, and Bachimont B. Mining defining contexts to help structuring differential ontologies. In Ibekwe-San Juan F, Condamines A and Cabré T, eds. Terminology, 2005. 11(1).To appear

[7] Zweigenbaum P, Bouaud J, Bachimont B, Charlet J, Séroussi B, and Boisvieux J-F. From text to knowledge: a unifying document-oriented view of analyzed medical language. Methods of Information in Medicine, 1998. 384-393.

Address for correspondence

Audrey BANEYX, INSERM U729 - Laboratoire SPIM, Faculté de Médecine Broussais-Hôtel-Dieu
15 rue de l'Ecole de médecine, 75006 Paris – France, Email: audrey.baneyx@spim.jussieu.fr, Mobile: 06 62 77 59 85

---

[4] The ICD-10 thesaurus is a medical and economic code for patients and allows analysing the activity of hospitals via diagnostic related groups.
[5] The French CCAM thesaurus is a complete and new classification of medical acts.
[6] http://www.biomath.jussieu.fr/Menelas/Ontologie/html/