Methodology to Build Medical Ontology from Textual Resources
Audrey Baneyx\textsuperscript{1} (MS), Jean Charlet\textsuperscript{1,2} (PhD), Marie-Christine Jaulent\textsuperscript{1} (PhD)
\textsuperscript{1} INSERM, U729, Paris, F-75006 France; and \textsuperscript{2}STIM, DSI/AP-HP, Paris, F-75014 France;

In the medical field, it is now established that the maintenance of unambiguous thesauri goes through ontologies. Our research task is to help pneumologists code acts and diagnoses with a software that represents medical knowledge through a domain ontology. In this paper, we describe our general methodology aimed at knowledge engineers in order to build various types of medical ontologies based on terminology extraction from texts. The hypothesis is to apply natural language processing tools to textual patient discharge summaries to develop the resources needed to build an ontology in pneumology. Results indicate that the joint use of distributional analysis and lexico-syntactic patterns performed satisfactorily for building such ontologies.

INTRODUCTION

For about ten 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 for the diagnoses codification. The French coding process is usually handmade by physicians thanks to medical specialty thesauri. These thesauri are built to help physicians code from common terminologies. It has been argued in a previous work, that 350 000 words is a minimum to build an ontology from texts. The hypothesis is to apply natural language processing tools to textual patient discharge summaries to develop the resources needed to build an ontology in pneumology. Results indicate that the joint use of distributional analysis and lexico-syntactic patterns performed satisfactorily for building such ontologies.

Objectives

We think that classification criteria used to design the hierarchical tree depend on purposes. We notice there is no ontology covering the pneumology 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 (cf. the OntoWeb Technical RoadMap, \url{http://babage.dia.fi.upm.es/ontoweb/wp1/OntoRoadMap/index.html}), 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. The methodology presented in this paper is based on differential semantics principles\textsuperscript{2}. Our main research hypothesis concerns the joint use of two methods to enrich the ontology building: \textit{i}) a method which consists in building terminological resources by distributional analysis\textsuperscript{8}, and \textit{ii}) a method based on semantic relationship recognition by the observation of corpus sequences showing the wanted relationship\textsuperscript{9}. First, we present the material and tools used in this work. The different steps of the methodology are then detailed. The results include statistical measures to evaluate the ontology, the coverage by the ontology of the specialty thesaurus and its use to assist the coding task. Finally, we conclude this paper by discussing the interest of our approach.

MATERIAL

In order to cover most activity of pneumology, we have gathered patient discharge summaries (corpus [PDS]) from six hospitals of the Assistance Publique-Hôpitaux in Paris-France. We ended up with a total of 1 038 patient discharge summaries. It has been shown in a previous work, that 350 000 words is a minimum to build an ontology from texts\textsuperscript{10}. The [PDS] corpus has about 417 000 words, the second corpus [BOOK] corresponding to a teaching book is about 823 000 words. We use SYNTEX-UPERY as a NLP tool\textsuperscript{11}. SYNTEX is a syntax analyser module based on the hypothesis of similar dependencies between terms having a close meaning. This module allows us to get relationships of syntactic dependencies between terms or syntagmas. At the end of the pro-
cess, we have a network of syntactic dependencies whose elements are the candidate terms that will be used to build the ontology. Then, the UP\textsc{e}RY module proceeds to the distributional analysis\textsuperscript{8}: it computes distributional proximities between candidate terms on the basis of shared syntactical contexts and exploits all the network data to cluster needed terms. Our material is composed by: 1) the network of candidate terms, 2) their contextual associations, 3) their links to the corpus. The results of analysis could be viewed with T\textsc{ermont}o, the data access and data process software interface. The Differential Ontology Editor (DOE, http://opales.ina.fr/public) allows us to build our ontology according to the Differential Semantics. This editor makes it possible to link manually multiple words from corpora to the same concept. Finally, the ontology is exported into OWL and formalized in description logics with The Protégé Ontology Editor (http://protege.stanford.edu/).

METHOD

We distinguish five successive steps in the method:

1) **Processing of resources by NLP tools**

The two [PDS] and [BOOK] corpora resources are collected in a format unexploitable by NLP tools. Files have been converted to text format, « cleaned », anonymised and morpho-syntactically analysed. We obtain an anonymous [PDS] corpus, and a didactical [BOOK] corpus, both in XML format. Next, the [PDS] corpus is processed by SYNTAX-UP\textsc{e}RY. The results of the analysis allow us to build basic elements – i.e. primitives – of the ontology. The second [BOOK] corpus is analysed to identify semantic relationships (hyperonymy, synonymy...) between candidate terms, thanks to previously defined lexico-syntactic patterns\textsuperscript{9}. The results help us control and enrich the hierarchy.

2) **Manual choice of candidate terms**

A candidate term (CT) is a noun phrase (NP) composed of 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. CT representative of pneumology are chosen within the results provided by SYNTAX-UP\textsc{e}RY from the [PDS] corpus in two steps:

1) We scan all the results provided by the syntactical analysis and choose to first study the CT 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 CT, we associate a classification criterion, available in T\textsc{ermont}o, that matches one of those axes, on a 1 to 6 scale: 1 (non pertinent terms, axis Other), 2 (reserved for terms which are modeled in the ontology), 3 (Signs), 4 (Pathologies), 5 (Diseases) and 6 (Treatments/examinations). For instance, we associate the classification criterion 6 with all CT on that axis - e.g. examination, doppler, radiography... At the beginning of the method, all the CT have a classification 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). Classification 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 classification criteria leaves 35% of CT.

2) The distributional analysis connects CT sharing the same expansions (Cf. figure 1). They are named scions in head and scions in expansion. This analysis is symmetrical, it also connects the expansions according to the CT they share (neighbours in head and neighbours in expansion). Scions in head bring information on what could be child-concepts or defined concepts. Scions in expansion bring information on the concept’s position in the hierarchy. Neighbours in head and in expansion let us constitute the groupings of CT semantically close. These groupings are a great help for the development of the ontological tree, for both horizontal and vertical axes.

3) **Application of differential principles**

In order to work that hierarchy out, it is convenient to organize the CT chosen by refining the differential principles that define them. For instance, the concept *FunctionalSign* and the concept *PhysicalSign* are sibling-concepts whose parent-concept is *ClinicalSign*. The principle of similarities with the parent-concept is to be a pathological manifestation visible without medical materials. The differential principle between sibling-concepts is related to the kind of examination practice in order to gather signs: a *functional sign* is a pathological manifestation felt by a patient, that only the patient can express and a *phys-
Figure 1: Contextual connections for the NP Effusion of pleura viewed with Termonto.

Physical sign is a pathological manifestation noticed by a physician at a medical cross-examination. The application of these four differential principles allows us to describe very precisely each concept in the hierarchy. The 4 axes (3, 4, 5, and 6) are refined that way. This part of the method is manual but the knowledge engineer is assisted in this task by the results provided by Syntex-upery. The analysis based on lexico-syntactic patterns recognition, processed on the [BOOK] corpus, gives clues for applying the differential principles. The lexico-syntactic patterns are representative of specific semantic relationships. These patterns are built around a marker, which is the lexical relationship indice like, kind of, for hyperonymy relationship. For instance, a pattern formed by DET NP, kind of NP allows the extraction of the lexical unit Meningitis, kind of pathology and gives a hyperonymy link between the nouns meningitis and pathology. We use 35 markers and 75 patterns. This method was presented in and experimented in several works. Lexico-syntactic patterns corresponding to hyperonymy build interesting parent-child pairs to control and enrich the hierarchical structure. To build differential ontologies, we apply this method to look for definition statements in the corpus (for instance: Dry cough is a sign for bronchitis.). The patterns used were developed by Malaisé et al. 9. The [BOOK] corpus is well structured and particularly suitable for this sort of research since it includes many definitions expressed in well structured language. The extracted lexical units are manually confirmed and the created hierarchies are available on DOE. Then, it is easy to compare the two terminological structures found: the hierarchy from the distributional analysis of [PDS] corpus, and the one from lexico-syntactic patterns analysis on the [BOOK] corpus. From that comparison, we figure out information for the enrichment and refinement of the ontology. At the end of that stage, we have a semantic normalization of the set of terms of the specialty and we have represented the hierarchy of primitive concepts and of relationships with DOE. The steps 4 and 5 of our methodology are processed by the Protégé Ontology Editor. For more information, the reader can refer to Alan Rector’s paper13.

RESULTS

The pneumology ontology

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 distributional analysis, the NP Chemotherapy of <noun> has the most neighbours in head, 28, and has 90 occurrences in the corpus. The NP <noun> of chemotherapy has the most neighbours in expansion, 52, and has the highest occurrence rate, 454. We can check the pertinence of the connections by groups of CT 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, thanks to 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/Cancerology/Navelbine, Cisplatine, Doxorubicine. . . The CT Antibiotherapy and Radiotherapy are also located under Treatment. 1050 concepts are directly identified from the results provided by Syntex-upery. The 150 other concepts are added by our expert from the French Pneumology Society. We have examined with special care the occurrences of medical signs or symptoms present in the [PDS] corpus. Their modelling is needed on the one hand to represent the patient, and on the other hand to detail and refine the expression of diagnoses to help coding. It is one of our major conceptual axes. According to our medical expert, we have relied on a teaching book about medical semiology to structure that knowledge axis. As for the presence of medical drugs in our ontology, the [PDS] corpus analysis revealed two different kinds of designations that we must model. Indeed, we have found both drugs designated from their commer-
The evaluation should test the completeness of the knowledge and the relationships contained in the ontology. Thus, we take into account the problem of International Nonproprietary Name versus trademarks. There are 22 names of classes, such as `PersonalHospitalRole`, `MedicalTools` or `ExaminationEnvironment`. These classes are top-classes, just under the high conceptual and philosophical level also named top-ontology. These classes are necessary to structure the concepts of the domain in the ontological tree. We complete the tree by connecting the top-ontology used in the MENELAS project ontology [16](http://estime.spim.jussieu.fr/Menelas/Ontologie/html/). MENELAS was about cardiology and our ontology is specific to pneumology. This work also checks if a conceptual common high level exists in the medical field, if it is possible to interoperate at this level. The top-ontology of MENELAS is slightly corrected and adapted. Its structure is preserved but we remove 23 concepts we do not need in our tree. Finally, our top-ontology is about 350 concepts, nearly 22.5% of the complete ontology of pneumology.

The definition recognition thanks to lexicosyntactic patterns on [BOOK] corpus allows extractions of 799 lexical units, 119 are confirmed, which represents 15% of extractions. The complementarity of the two hierarchies is interesting as the result of different methods applied to different kinds of textual corpora. The divergence of results analysis gives a good critical point of view to the knowledge engineer for modelling the specialty.

Our ontology includes today 1200 primitive concepts stemming from the first analysis of the CT. The building stages 1 and 2 are iterative, and we can quickly increase the hierarchy by examining the CT which appear less than 12 times in the [PDS] corpus. At the end of each phase, a knowledge engineer validates the results and helps us enrich the representation. The ontology must be evaluated in terms of both quality and coverage. The complete ontology contains 1550 concepts, 412 relationships and the depth of the tree is about 18.

The coverage of the specialty thesaurus

The evaluation should test the completeness of the ontology compared to the French specialty thesaurus used as reference by pneumologists to code their activity. To estimate the coverage, we have to check the possibility of building a conceptual representation of wording of medical knowledge by combining the primitive concepts and the relationships contained in the ontology. 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. For example, the part of the thesaurus about obstructive chronic broncho-pneumonopathy (a group of diseases which involves a limitation of the passage of the air in lungs) contains 17 labels such as `Simple chronic bronchitis`. The level of coverage of this chapter is 92 implying 38 concepts and 16 relationships. The 8% of non coverage are due to vagueness. For example, we are able to represent `Simple chronic bronchitis` but not precisely `Superinfection of chronic bronchitis` because the ontology contains the concept `Infection` but not contains the concept `Supernfection`.

The ontology allows us to identify a number of labels much more significant than those found in the thesaurus, for example by linking every concept of the class `Sign` to every concept of the class `AnatomicalObject`. The methodology gives us the possibility to easily add new concepts in the ontology. The differential principles guarantees the addition of new concepts in the right place in the ontological hierarchy.

Coding of patient discharge summaries

The first purpose of the ontology is to help pneumologists code acts and diagnoses. To satisfy this objective, the ontology of pneumology is converted according to a specific model into a Mysql database which is the heart of our coding tool. The labels of the specialty thesaurus are linked to corresponding defined concepts. They establish a link between the terms of the language and the concepts of the ontology. The coding tool allows the pneumologist to open a patient discharge summary in the interface and offers the following functionalities: 1) a description of the medical acts and diagnoses in the form of conceptual graphs and 2) the proper medical-economic French code. This software tries to represent the patient’s data with all the concepts and relations the pneumologist finds relevant. For the description of diagnoses and acts, we obtain a recall of 80% and a precision of 87%. This evaluation is made manually on 500 PDS chosen randomly. By analyzing PDS, we could envisage the relevant terms which must appear and their number. For the medical-economic coding, we obtain a recall of 70% and a precision of 85%. This evaluation is made on 100 PDS which were previously coded manually by pneumologists.
DISCUSSION AND CONCLUSION

We have presented a range of methodological principles for the building of differential medical ontologies. This work demonstrates the usefulness of NLP tools in modelling a domain from textual resources. The actual coverage of the specialty thesaurus, for the Obstructive Chronic Bronchopneumonopathy chapter, the Infectious Pathology chapter and the Tumor chapter, validates the choice of the corpora. The textual corpora, like the ones we are using, refer to specific and consensual vocabulary well diffused and shared by medical corporation. This is a guarantee of reliability and stability for our modelling. We also have shown the joint use of two methods, distributional analysis and use of lexico-syntactic patterns, each of them adapted to a kind of corpus. The complementarity of the two hierarchies is interesting as the result of different methods applied on different kinds of textual corpora. The divergence of results analysis is rich in information and gives a good critical point of view to the knowledge engineer on the way of modelling the domain. We also found limits in the comparison of the two methods: bringing together the hierarchies with automatic tools implies, on the one hand, more sophisticated pairing technicals and on the other hand, precise the lexico-syntactic patterns by adapting them specifically for the medical field.

Research around building medical ontologies is more and more concerned by a formal representation of the patient data. This representation has to be formal for two reasons. First, a high interoperability should rely on the representation in order to be sure that generated knowledge in health is useful and accessible for the medical analysis. Second, it has to allow the coding of acts and pathologies for medico-economic purposes. From the point of view of medical analysis, physicians are working with are interested in formal representations of knowledge from patients’ records to be able to perform any epidemiological studies from patient data. To do that, an ontological representation is essential. Then, it seems important to offer a precise, complete and specific ontology to the modeled field. Finally, we develop specific ontologies, connected to the thesaurus of the domain if any. From that point of view, our work is close to the GALEN team (http://www.opengalen.org/).

REFERENCES


