Terminology extraction from text to build an ontology in surgical intensive care

Sophie Le Moigno[1] and **Jean Charlet**[2] and **Didier Bourigault**[3] and **Patrice Degoulet**[1] and **Marie-Christine Jaulent**[1]

Abstract. In the medical field, and in the many specialities which it is made of, it is now established that the maintenance of unambiguous terminologies or the comparison and aggregation of different terminologies goes through the building of formal specialized clinical terminologies, the ontologies.

Here, we describe the building of an ontology in surgical intensive care. This work is based on the utilization Natural Language Processing tools, corpus analyser and a distributional analysis tool. Especially we have tested the possibility for an expert to build a sizeable ontology in a reasonable time. The quality of the ontology has been evaluated according to its capacity to cover the CIM-10 terminology in the correspondant speciality.

1 INTRODUCTION

Surgical intensive care is a medical domain specialized in the treatment of post-operative complications and in traumatology. This specialty includes various pathologies and acts that are classified in thesaurus to help the physician to code his/her activity. Variability in coding is a well-known problem due in part to the ambiguity of thesaurus [4]. Some work has been done on automatic coding tools to reduce mistakes and variability of coding [6]. However, it has been argued in the literature that the task can not be achieved without a formal organization of the concepts of the domain within an ontology [2]. One important reason for that, among others, is the difficulty, for computer scientists, to develop and maintain complete and non ambiguous thesaurus, to compare or merge several thesaurus [14][17][10].

The objective of the present work was to achieve an ontology of the surgical intensive care domain. In this context, we considered textual reports as the main source of information and Natural Language Processing (NLP) tools were used to build a sturdy ontology.

In this paper, we present at first current works on ontology construction and more particularly from texts of the domain. Our methodology, based on the SYNTEX and UPERY corpus analysers [7] [7], is then detailed. The results include statistical results, such as the number of concepts or the development time provided by the interpretation of the NLP tools outcomes, the ontology itself described as an " is-a " hierarchy of concepts and a hierarchy of relations. Finally, the coverage of the thesaurus by the ontology is discussed.

2 BACKGROUND

Among existing medical concepts organizations one can found:

The SNOMED-RT, under development, uses existing axes of the multi-axial nomenclature of the SNOMED [13] to define the concepts. Every diagnosis is linked explicitly to morphological, anatomical and etiological properties.

The MAOUSSC project is a multiaxial formalism for the representation of surgical concepts in the context of automatic coding [12].

GALEN (General Architecture for Language and Nomenclatures) is a system dedicated to the development of ontology in all medical domains including surgical procedures [3][15]. In this domain, this nomenclature is designed to answer multi-lingual coding objectives and to compensate for the deficits of nomenclatures and existing classifications.

MENELAS is a Natural Language Processing (NLP) system inside which a French ontology has been developed in the domain of coronaropathies [19]. It has been elaborated manually from the analysis of texts of the domain [2] some years ago at a time where Natural Language Processing (NLP) tools were not yet of frequent use.

Nevertheless, nowadays, no ontology currently exists in the domain of surgical intensive care.

Organization of knowledge is a difficult and time consuming task. The knowledge engineering community distinguishes two classes of methods to help to perform this task.

The downward method (KADS) favors specific concepts definition from generic models [16] and the ascending method (KOD) uses directly the meaning of words to organize the knowledge [18]. This method includes texts analysis but NLP tools are not often used.

The differential semantics describe the concepts using their resemblance as well as their dissemblance. In this sense, a method has been described more lately by Bachimont [2]. It is based entirely on texts analysis and give a lot of place to the respective roles of the expert and the texts in the conception of the ontology. This method is close to ascending methods.

NLP tools, based on the differential approach, have been developed to allow the definition of classes of concepts and relations using a distributional analysis of the contexts of the terms in a given corpus [11]. These tools already contributed extensively to the construction of ontologies [1]. Different experimentations showed that great attention should guide the choice of the corpus and tools [9].

Bouaud and al. brought up the advantages, in the linguistic approach, of using syntactic analysis tools (ZELLIG) to model domain terms, rather than using no tool (MENELAS) [5].

In this article, we present how NLP tools have been useful for (1) reaching a mass of information or knowledge sometimes very

[1] SPIM, UFR Broussais-Hotel-Dieu, Paris , France

[2] Mission recherche STIM, DPA/DSI/AP-HP & Dép. Biomath, Univ. Paris 6, France

[3] Équipe de recherche en syntaxe et sémantique, CNRS, Univ. Toulouse II, France

important and inaccessible manually [1] and (2) associating the different previously described approaches in order to ease and minimize the expert's task as well as to promote the linguist's one.

3 MATERIAL

In this work, we took textual reports (CRH) from surgical intensive care domain. The corpus is made up of CRH from seven surgical intensive care units in "Ile-de- France". Units were selected so that the distribution covers a large set of the surgical intensive care activities (traumatology, cardiovascular, general surgery, neurosurgery and neuro-traumatology and obstetrics). 800 CRHs have been collected initially.

A first analysis was done to determine the optimal number of CRH in order to cover the most concepts of the domain as possible. According to G.K Zipf, when words of a text are classified by decreasing frequency, then the frequency of a word is inversely proportional to its rank [20]. The frequency of the second word is then the half of the first one and the frequency of the third is its third. We applied the Zipf law on several subset of our corpus (i.e. respectively 200, 400, 600 and 800 CRHs). From this preliminary study, 600 CRHs have been kept in the final corpus and restricted to the traumatology field which was predominantly represented in the initial CRHs.

In addition, the goal standard used for the evaluation of the ontology is a thesaurus of the speciality written in 1999 <www.sfar.org>. Traumatology corresponds to a specific chapter in this thesaurus.

Before using any NLP tool, the corpus has been tagged in textual units. The tagging procedure allows to affect a specific flag to different paragraph types of the original text. In the tag example <*AB23-HDM>, "AB" indicates which unit the CRH come from, "23" corresponds to the number of the CRH in this unit and "HDM" indicates that the paragraph is about "History of the Illness". This procedure is important since it allows to recover every term in the initial text after analysis.

4 METHODS

4.1 The *SYNTEX* analysis

The ontology was built using the results yielded by the SYNTEX software and the UPERY module [7]. SYNTEX is corpus syntactic analyser. It performs a syntactic analysis of the sentences of the corpus, and yields a dependency network of words and syntagms. A verb syntagm (resp. noun or adjective syntagm) is a group of words with a head being a verb (resp. a noun or an adjective). For instance the verb syntagm "reveal a bone lesion" is composed of a head – reveal - and an expansion (a noun syntagm) – a bone lesion -. Each syntagm in the network is connected to its head (H) and to its expansion(s) (E), the link being labeled by the name of the dependency relation (R). The UPERY distributional analysis module compute semantic proximities between words or syntagms in the network using the notion of shared syntactic contexts. The approach is based on the distributitional analysis principle [11]. Every link (H, R, E) is represented as a couple (Context, Term) where the Context is the concatenation of the Head and the Relation and where the Term is the Expansion. Given a context C, the quantity Prod(C) is defined by the number of Terms (words or syntagms) associated to C. Conversely, given a term T, the quantity Prod(T) is defined by the number of Contexts in which T occurs.

The distributive analysis brings closer Terms that share a large number of Contexts. It also brings closer Contexts that share same Terms. Three measures are defined to calculate proximities between two entities (Terms or Contexts).

- *The "a" coefficient*. Given two Terms, "a" is defined by the number of syntactic Contexts shared by the two Terms.
- *The coefficient "prox"* measures the shared context productivity. It is defined by the number of Terms that appears in this Context.
- *The coefficients j1 and j2* characterizes the proximity between two terms and measures the number of contexts that every Term have of its own. Given two terms T_1 and T_2,

  $j_1 = a / prod(T_1)$

  $j_2 = a / prod(T_2)$

The UPERY module computes these coefficients for every couple of Terms and every couple of Contexts. Only relevant couples (for which coefficients are above some thresholds) will be considered for the construction of the ontology. In our domain, the thresholds have been defined heuristically and provide the following constraints: (1) "a > 3", "prox > 1" and "$j_1 > 0.5$or $j_2 > 0.5$"which means that one of the two terms must share at least the half of its contexts with the other term.

4.2 The construction methodology

In this work, the results provided by SYNTEX and UPERY were the basis for the construction of the ontology by a physician expert of the domain (not a computer scientist and not a linguist). The expert analyzed couples of terms (noun, adjective) and couples of contexts (verb, noun) presenting high proximity coefficients. The work contributed to establish a methodology allowing to extract, as fast as possible, from an enormous amount of information, the terms and the most representative associations of terms of the domain. This methodology contains two phases. In a first time, a method is proposed to extract large classes of concepts by using the results of the distributive analysis. In a second time, classification of the concepts was refined coming back to the Head and Expansion analysis in the terminology network.

Phase 1 : spadework on classes

1. Analysis of couples made of close verb contexts in order to come out with different types of verbs. For the domain, 2 classes of verbs were extracted: state verbs (to present, to show,…) and action verbs (to achieve, to indicate,…).
2. Analysis of expansions (subject or complement) of the previous verbs in order to extract others large classes. For example, an action verb will have a therapeutic or diagnostic action for direct complement while a state verb will have a pathological or physiological state for direct complement. Names of states and names of actions have types of localization as expansions. The creation of an ANATOMY class follows this observation.
3. Identification of "key couples" that highlight the creation of new classes by analysis their expansions. For instance, the DRUG class derived from the analysis of the expansions of the key couple "administration – introduction". The other expansions, that do not belong to the new class, are assigned one by one to existing classes if possible. When it is not

possible, a new class is created.

4. Completion of classes. A class is composed of terms for which expansions are close, although not identical. Expansions that differ from one couple to another are analysed to complete existing classes.
5. The procedure must be followed until a certain limit. This limit is reached when no new class is created by "key couple" analysis in a reasonable work time.

Phase 2. Refining of classes.

At the end of the previous phase, in order to construct the ontological tree, it is sufficient for the expert to take terms of the new classes. The objective of this phase is to analyse in the terminology network the Head and Expansion descendants of the terms or the noun syntagms they compose. This phase enriches classes (ex: extraction of the concept "syndrome" that has a few shared contexts with other terms) and to establish relations between two concepts of a same syntagm already present in the hierarchical tree but in different classes.

5 RESULTS

5.1 The surgical intensive care ontology

The distributive analysis extracted 650 couples of verb contexts. The couple with closest coefficients is "show - recover". The number of terms shared by these two contexts is equal to 102. It means that, in the corpus, the verbs "to show" and "to recover" have 102 direct object complements in common.

The distributive analysis extracted 12052 couples of noun terms. The couple with closest coefficients is "fracture– lesion". The number of verb context shared by these two terms is equal to 10. It means that, in the corpus, the nouns "fracture" and "lesion" appear (at least once each) as direct object complement of those verb contexts that are "to show", "to evoke", "to have" etc.

The distributive analysis extracted 38620 couples of noun contexts. The couple with closest coefficients is « niveau - fracture ». The terms shared by these contexts are nouns or adjectives relating to an anatomical (basin, cervical) or geometric (bilateral, right) location.

From the successive analysis of verb and noun contexts and of couples of terms (noun and adjective), several "key couples" have been extracted:

1. The class (transfer, arrival, departure, hospitalisation) is associated, as contexts, to medical concepts (i.e. orthopaedics), to medical units concepts (USPI) and action concepts.
2. The class of grading importance (important, severe, moderate, light, etc.) is associated, as terms, to the local pathological states (lesions, symptoms) or dysfunction (insufficiency).
3. The class of assessment level (elevated, correct, good, satisfactory) is associated, as terms, to measure concepts (diuresis, etc.) and to diagnostic acts.
4. The geometric location class (right, left, previous, posterior, etc.) is associated, as terms, to the anatomical objects (i.e. lung), local pathological states (i.e. extravasations) and to the pathological processes (i.e. traumatism).
5. The pathological state class is associated, as terms, to the evolution concepts (i.e. improvement, aggravation, etc.).
6. The global actions classes (administration, introduction, cessation, weaning) are associated, as contexts, to the drug products and administration methods (withdrawal, rise) are associated, as contexts, to the instruments.
7. The circumstance class (emergency) is associated to actions.

At the end of the first phase, 984 concepts are extracted. 819 concepts are directly identified from the terms candidates provided by SYNTEX. The 165 other concepts (added or modified by the expert) are Names of classes. This result has been obtained in 120 hours by an expert of the domain.

The second phase of the analysis ended up with the definitive ontology. This ontology contains 2114 concepts (class concepts, primitive concepts, definite concepts and concepts of relation) and 185 relations. 69 concepts are concepts of relation. 1730 concepts are directly derived from term candidates provided by SYNTEX and 385 concepts (22%) have been added or modified by the expert. Among these 1730 concepts, 988 (57%) are nouns, 633 (36%) are noun syntagms and 109 (6%) are adjectives. 60 additional hours were necessary to obtain this result. During this phase, some classes underwent some modifications but the initial organization was overall preserved.

5.2 Coverage of the thesaurus

On the 658 pathologies labels in the thesaurus, 400 (60%) are covered by the ontology. In the traumatology domain, the level of coverage is of 100%. An example of coverage of the thesaurus is presented in figure 1 ("Intracerebral traumatic lesion with prolonged coma" is a label present in the thesaurus).

Other domains are less covered by the ontology like the dermatology (36%) or the hematology (12,5%) domains.

Intracerebral traumatic lesion with prolonged coma

[LESION] (ETIOLOGY : caused_by)
[EXTERNAL_PATHOLOGIC_PROCEEDING:traumatism
]
(LOCATION:inside)
[BONE :skull]
(EXPRESSION : to show up)
[OBJECTIVE_ SIGN : coma]
(REL_EVOLUTION : develop_ so as to)
[VALUE_NIV_EVOLUTION : prolonged]

FIGURE 1. Example: Cover of the thesaurus

In the ontology, every lesion or process can be linked to every anatomical element. By linking every concept of the class LESION to every concept of the class ANATOMY, it is possible to identify a number of concepts much more important than what can be found in the thesaurus. Conversely, some concepts present in the thesaurus are absent of the ontology. The possibility to add easily new concepts to the ontology shows the sturdiness of the approach. The following example illustrates the procedure to add a new concept in the ontology.

Example: Cardiac septum malformation

As the term "septum" can be associated to the term "nasal", it is easily positioned in the SITUATION_DANS_OBJET class without modifying the ontology structure.

5.2 Coding of reports

In order to illustrate the ontology performances for the coding activity, six CRHs, different from the corpus, have been coded manually by using labels present in the ontology. The same case were coded by a physician in a classical way using the thesaurus. 26 labels are used for the coding with the ontology against 21 by the physician. Among them, 11 labels agree, 10 are coded by the physician and not by the ontology and, in the opposite, 15 are non coded by the physician are while recovered by the ontology. Moreover, we found 9 of the 10 missing labels from the classic coding were not expressed in the text. Differences in coding are illustrated for one CRH in table 1.

Table 1. Example of differences between classic coding and ontology coding

Classic coding	**By ontology Coding**
Limb open fracture T 131	
Rib fracture S22.3	**Multiple rib Fractures S22.4**
scapular arch fracture S42.9	**scapular arch fracture part SAI S42.9**
Head wound S019	
Psychiatric disorder F99	**Psychiatric disorders and démentia SAI F99**
Agitation R45.1	**Agitation R45.1**
Bone thorax Fracture S22.90	
	Skull and face bones Fractures, part SAI S02.9
	suicide attempt X84.9

6 DISCUSSION AND CONCLUSION

This work demonstrates the usefulness of NLP tools in modelling a domain from texts. Indeed, some previous works showed the difficulty to develop an ontology (MENELAS [19]). In the current work, the expert need only 200 hours to construct this ontology. The corpus contains 600 CRHs corresponding to a limit we established, from which new concepts are under- represented. The actual coverage of the thesaurus, for the chapter “traumatology”, does not put into question the choice of the initial corpus content.

Our methodology based on “key couple” identification allows to extract quickly classes and relations. In this methodology, the “key couple” is not always the one that provides the most shared contexts. However, the key couple is often the first couple recovered in the list of terms having the same shared contexts. One interesting extension of the SYNTEX software would be to extract automatically "key couples " and missing expansions.

The role of the expert in the construction of the ontology is crucial and the proposed methodology guides his different abilities for :

- the choice of a “key couple”. For instance, according to his knowledge, the expert may decide that the couple (hemothorax - hemoperitoine) is more interesting than the couple (hematoma - syndrome) although the second one is the first of the list,
- the distinction between terms of a same class and synonymous terms,
- the choice of a class for special terms. For instance, "pain" is a clinical symptom often recovered in the lesion list because of its association with an anatomical location and is not specifically associated to the other symptom concepts. Only the expert can define this concept as " symptom " and non as “ lesion”.

The objectives of this work are not very different from those of MENELAS and get close to those of the GALEN team. The originality of the work is in the method based on NLP tools. The results show the necessity to get a robust organization of concepts in order to obtain a robust thesaurus. An immediate consequence would be to facilitate the maintenance of large size thesaurus.

Further work has to be done to reach the level of an helping coding system. In particular a representation formalism is required (logical description or conceptual graph) to design a computerized system to go from a code to another. However the accomplished work is a mandatory stage toward an helping coding system.

REFERENCES

[1] Assadi H, Bourigault D. Analyses syntaxique et statistique pour la construction d’ontologies à partir de texts. In : Ingénierie des connaissances. J.Charlet, M.Zacklad, G.Kassel, D.Bourigault (eds). Eyrolles:Paris - Collection technique et scientifique des télécommunications, 2000; pp243-255.

[2] Bouaud J, Bachimont B. , Charlet J, Zweigenbaum P. Methodological principles for structuring an “ontology”. *In:* IJCAI'95 Workshop on “Basic Ontological Issues in Knowledge Sharing”, 1995; pp95-148.

[3] Baud RH, Lovis C, Rassinoux AM, Scherrer JR. Alternative way for knowledge collection , indexing and robust language retrieval. *Meth Inform Med* 1998; 37: 315-26.

[4] Berthelsen CL. Evaluation of coding data quality of the HCUP National Inpatient Sample. *Top Health Inf Manage* 2000 Nov;21(2):10-23

[5] Nazarenko A, Zweigenbaum P, Habert B, Bouaud J,. Corpus-based Extension of a Terminological Semantic Lexicon. *In* : Recent Advances in Computational Terminology. D Bourigault, C Jacquemin, and Marie-Claude L'Homme (eds). John Benjamins: Amsterdam, 2000; pp.

[6] Bouchet C, Empereur F, Kohler F. Evaluating a computerized tool for coding patient information. *Proc AMIA Symp* 1998;185-9.

[7] Bourigault D. & Fabre C., Approche linguistique pour l’analyse syntaxique de corpus, *Cahiers de Grammaires*, n° 25, Université Toulouse - Le Mirail, 2000; pp. 131-151

[8] Bourigault D., Analyse distributionnelle étendue, Actes de la 9e conférence annuelle sur le traitement automatique des langues, 24-27 juin 2002, Nancy

[9] Condamines A. Aide à l’acquisition de connaissances par l’étude de la terminologie. *In : acquisition et Ingénierie des connaissances, tendances actuelles* ; Aussenac-Gilles N, Laublet P, Reynaud C (eds) Cépadues-Editions : Toulouse. 1996 ; pp247-265.

[10] Dolin RH, Spackman K, Abilla A, Correia C, Goldberg B, Konicek D, Lukoff J, Lundberg CB. The SNOMED RT Procedure Model. Proc AMIA Symp. 2001;2001:139-143.

[11] Harris Z, Gottfried M, Ryckman T, Mattyck P, Daladier A, Harris T, Harris S. The form of information in science,

analysis of immunology sublanguage. In Boston Studies in the philosophy of science 1989; volume 104. Kluwer academic publichers, Dordrecht, Pays-bas.

[12] Levêque J, Burgun A, Foucher F, Levêque JM, Grall JY, Le Beux P. Analysis of medical catalog terms using the MAOUSSC model: application to Gyncecology-Obstetrics. *J Gynecol Obstet Biol Reprod* (Paris). 1998 Nov;27(7):676-82.

[13] Levy DH, Dolin RH, Mattison JE, Spackman KA, Cammpbell KE. Computer-facilitated collaboration : experiences building SNOMED-RT. *Proc AMIA Symp* 1998; 870-4.

[14] Rector AL, Solomon WD, Nowlan WA, Rush TW, Zanstra PE, Claassen WM. A Terminology Server for medical language and medical information systems. Methods Inf Med. 1995 Mar;34(1-2):147-57.

[15] Rodrigues J-M, Trombert-Paviot B, Rector A, Baud R, Clavel L, Abrial V, Idir H, Very J-M. Galen, il existe quelque chose après les mots : leur signification et au delà le savoir médical. *Innovation Stratégique en Information de Santé (ISIS)* 1999; 2-3:48-62.

[16] Schreiber A, Wielinga B, Breuker J. KADS: a principled approach to knowledge-based system development. In *Knowledge-based Book Series* 1993. Academic Press, London: volume 11.

[17] Spackman KA, Campbell KE, Cote RA. SNOMED RT: a reference terminology for health care. Proc AMIA Annu Fall Symp. 1997;:640-4.

[18] Vogel C. Génie cognitif Paris : Masson 1988

[19] Zweigenbaum P. Encoder l'information médicale: des terminologies aux systèmes de représentation des connaissances. *Innovation Stratégique en Information de Santé (ISIS)* 1999; 2-3:27-47.

[20] users.info.unicaen.fr/~giguet/java/zipf.html. consulté le 5/06/01