Model-based Semantic Dictionaries for Medical Language Understanding

Anne-Marie Rassinoux, Ph.D.[1], Robert H. Baud, Ph.D.[1], Patrick Ruch, M.S.[1]
Béatrice Trombert-Paviot, M.D.[2], Jean-Marie Rodrigues, M.D.[2]
[1]Medical Informatics Division, University Hospital of Geneva, Switzerland
[2]Department of Public Health and Medical Informatics, University of Saint Etienne, France

Semantic dictionaries are emerging as a major cornerstone towards achieving sound natural language understanding. Indeed, they constitute the main bridge between words and conceptual entities that reflect their meanings. Nowadays, more and more wide-coverage lexical dictionaries are electronically available in the public domain. However, associating a semantic content with lexical entries is not a straightforward task as it is subordinate to the existence of a fine-grained concept model of the treated domain.

This paper presents the benefits and pitfalls in building and maintaining multilingual dictionaries, the semantics of which is directly established on an existing concept model. Concrete cases, handled through the GALEN-IN-USE project, illustrate the use of such semantic dictionaries for the analysis and generation of multilingual surgical procedures.

INTRODUCTION

The success of medical language understanding (MLU) is largely dependent on the existence of wide-coverage, fine-grained semantic dictionaries. In analysis, each form encountered in free-text sentences should be checked against the dictionary's lexical entries. These forms can be words, possibly decomposed in terms of morphosemantic constituents, or multi-word expressions. Syntactic and semantic information, defined in the dictionary for each basic form, will serve to correctly analyze sentences and build their corresponding language-independent representation. In generation, semantic dictionaries are examined from the semantic information that precisely constitutes the input structure given to the generator. Therefore, they render all the words, together with their syntactic categories, that are permitted to express a specific meaning using appropriate language structures. In both cases, semantic dictionaries bridge a gap between words and their associated meanings in a specific domain, thus allowing reasoning to be triggered from the domain structure.

The acquisition and representation of lexical and semantic knowledge are fundamental issues in the medical informatics community[1]. The linguist's ability together with the domain engineer's knowledge are required to respectively extract lexical entities from medical corpora and represent their meaning in a consistent manner. With the electronic availability of large medical corpora (such as patient narratives) as well as controlled medical vocabularies[2], lexical and statistical methods have been largely exploited to automatically acquire linguistic resources[3, 4, 5, 6]. While results have proved to be significant for acquiring lexical entries, the automatic acquisition of the semantic content of words remains limited to capturing structural relations among a set of related words (i.e. co-occurrence patterns).

An additional approach for building semantic dictionaries consists of adding lexical entries directly to an existing concept model describing the semantics of the treated domain. Such a model-based solution presents interesting features for building dictionaries for MLU. The quality of the lexical annotations as well as the constraints associated with the model structure, constitute the main focus of this paper. After reviewing the characteristics of model-based semantic dictionaries, their use within the analysis and generation processes is discussed and directions for further work are finally exposed.

BACKGROUND

The large volume and diversity of medical knowledge make the conceptual modeling task difficult and labor-intensive. Once built, however, concept models provide language-independent and structured domain knowledge upon which various multilingual applications can be set up.

Using the GALEN Model for MLU

The GALEN project has developed a common reference model for medical concepts (the so-called CORE Model) that is supported by formal language for concept representation (so-called GRAIL)[7]. The current phase of the project, renamed GALEN-IN-USE, applies these tools to assist in the collaborative construction and maintenance of surgical procedure classifications. Our contribution to this task was to deliver a multilingual natural language toolkit, including both an analyzer and generator of surgical

1091—8280/99/$5.00

procedures[8]. Exploiting the CORE model for MLU consisted of adding a linguistic layer to the two main semantic components of the model, namely the typology of concepts and the semantic rules that govern the sensible combinations of these concepts. The first linguistic task relates to the language annotation process for concepts, resulting in the creation of multilingual dictionaries. The second linguistic task concerns the annotation process for semantic rules. This consists of clarifying syntactic structures that are commonly used in a given language to support the expression of the relationships occurring between conceptual entities.

The rest of this paper will focus on the former task dealing with the creation and tuning of such semantic dictionaries for different languages. Numerous examples, translated in English for the sake of clarity, stem from the elaboration of a new French coding system for surgical procedures called CCAM[9], which is currently being developed by the University of Saint Etienne.

Multilingual Annotations of Concept Model

Traditional approaches for building dictionaries start by grasping all the lexical entries (subsequently reduced to their basic forms) found in a given domain. Afterwards, 'computable' information describing the associated syntactic and semantic knowledge is depicted for each basic form. When relying on a concept model, the approach is rather different. Indeed, building dictionaries corresponds to annotating concepts with words that best express their underlying meaning (see the annotation of the concept *cl_Joint* in Figure 1).

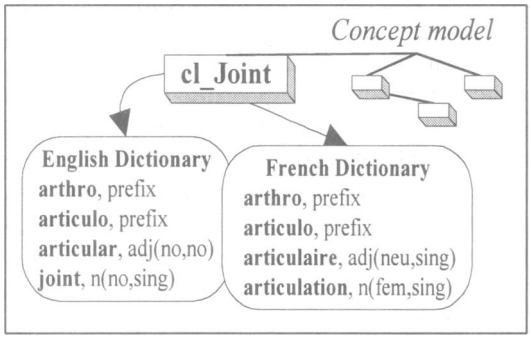

Figure 1 - Building concept-based dictionaries

In this way, it is only permitted to attach words to existing concepts in the model. This has strong repercussions on the size of dictionaries that are then regulated by the model coverage. However, such a constraint is beneficial for maintaining annotations in various languages, as concepts act as an interlingua or language-independent representation. This allows annotations in a specific language to be precisely grasped for a given concept, working by analogy with what has already been done in other languages. Finally, matching concepts directly with words raises new issues, such as: - Which concepts need to be annotated? - How accurate must the annotation be? - What kind of artifice must be set up to conciliate language idiosyncrasies with the model style? These questions are discussed in the following sections.

FEEDING MODEL-BASED SEMANTIC DICTIONARIES

The content of dictionaries is a critical factor for setting up large MLU applications. The analysis process requires a good coverage of the treated domain in order to successfully recognize and interpret 'any' or at least 'most of the' input narrative. Such coverage affects both the lexical variations (i.e. enumerating all syntactic categories that denote a concept) and semantic variations (i.e. taking into account the various meanings for one particular lexical entry). Regarding the generation process, it is less exigent as it expects the dictionary to return at least one annotation for a concept according to a predefined syntactic category. However, additional annotations would ensure a richer expressiveness of the generated sentences.

Meaningful versus Theoretical Concepts

Linguistic annotation of concepts defined in the domain model must be performed selectively in order to give priority to the annotation of concepts that embed practical medical meanings. Those concepts must be distinguished from theoretical concepts that are only there for the purposes of the modeling process. The latter include concepts that serve to organize the high level of the domain knowledge (e.g. the GALEN concepts *cl_GeneralisedSubstance* or *cl_GeneralisedStructure*), or that denote abstract or arbitrary categories useful to assist the modelers with the structure of the model (e.g. the GALEN concepts *cl_NAMEDPathologicalProcess* or *cl_ArbitraryBodyConstruct*). Taking into account the modeling style, an automatic annotation process for the GALEN model has been successfully carried out for creating an English dictionary. Indeed, each English knowledge name, effectively used to label a relevant concept in a form readily understandable by human beings, was chosen as a potential English annotation. This assumes that modelers choose knowledge names that are unambiguous with a well-accepted medical usage. For other languages, annotations were entered manually, working by analogy with the English version.

Compositionality of Language and Model

While considering relevant domain concepts to be annotated, it is also of paramount importance (especially for the analysis process) to grasp the various linguistic expressions that are permitted to express a specific meaning. However, enumerating all lexical variants can rapidly become unbounded and time-consuming. For example, the English expressions *arthritis*, *articular inflammation*, or *inflammation of joints* are valid annotations for the GALEN concept *cl_Arthritis*. In order to avoid specifying all these expressions, one solution is to exploit the compositionality of the medical language[10] as well as the combinatorial aspect of the compositional model approach[7]. On the linguistic side, this relates to the vocabulary conciseness allowing medical compound word forms derived from Greek and Latin to be used. On the conceptual side, this mostly deals with the granularity of the representation allowing composite concepts to be defined through more basic concepts.

Both situations are strongly interconnected, as the semantic interpretation of compound words that are built from morphosemantic constituents, can be regarded as semantic definitions that are expressed and maintained at the conceptual level. Given the previous example, the concept *cl_Arthritis* (i.e. *arthro* + *itis*) is maintained in the domain model through the following definition:

[*cl_InflammatoryProcess*]-
-> (*rel_actsSpecificallyOn*) -> [*cl_Joint*]\\.

MLU tools, making use of the annotations defined for the basic concepts *cl_InflammatoryProcess* and *cl_Joint* can then procedurally handle all the above mentioned lexical variants. Moreover, the agreement and expansion of multi-word expressions that cause problems when set as lexical entries, become manageable through the compositional approach. This allows expressions such as *inflammation of several knee joints* to be correctly handled.

However, sometimes composite concepts need to be directly annotated with concise words, in order to be suitable for linguistic generation. This is the case for the composite concept *cl_Fracture* that is defined by:

[*cl_BodyStructure*]-
->(rel_*hasUniqueAssociatedProcess*)
->[*cl_FracturingProcess*]
->(*rel_hasPathologicalStatus*)->[*cl_pathological*]\\.

Deciding whether to annotate or not a composite concept becomes strongly dependent on the nature of information described in the associated definition as well as on the availability of specific annotations in the treated language.

Accuracy of the Annotation Process

As the chosen approach for building semantic dictionaries is based on the existence of a concept model, the precision of the annotation process is becoming strongly dependent on the granularity of the model. First, only named concepts can be annotated, thus excluding the anonymous GALEN concepts that are only there for use within other definitions. Second, annotations must reflect, as much as possible, the intended meaning of concepts. Let us consider the following excerpts of French CCAM rubrics[9] as modeled within the GALEN model:

(1) Rubric U392: *Epilation endo-uréthrale…* (i.e. *Endo-urethral epilation…*)
[*cl_Removing*]-
->(*rel_actsSpecificallyOn*)-> [*cl_Hair*]-
->(*rel_isContainedIn*)->[*cl_Urethra*]\\.

(2) Rubric S027: *Avulsion d'une canine maxillaire…* (i.e. *Avulsion of a maxillar canine…*)
[*cl_Removing*]-
->(*rel_actsSpecificallyOn*)-> [*cl_Canine*]-
-> (*rel_isSolidRegionOf*)->[*cl_Maxilla*]\\.

(3) Rubric S208: *Dépose d'un biomatériau facial…* (i.e. *Depose of facial material…*)
[*cl_Removing*]-
->(*rel_actsSpecificallyOn*)-> [*cl_Material*]-
-> (*rel_involves*)-> [*cl_Face*]\\.

In the above examples, taken from the urology and stomatology chapters of the CCAM classification, the particular concept *cl_Removing* has been chosen for semantically representing the English words *epilation*, *avulsion*, and *depose*. However, they cannot be considered as synonymous as their linguistic form sensibly carries particular significance. Considering these shades of meaning as equivalent annotations for the concept *cl_Removing* would be awkward, especially for the generation process, which expects the dictionaries to provide words that closely reflect the meaning carried out by concepts. The solution chosen was to explicitly create and annotate three new composite concepts as descendants of *cl_Removing*. Those include the removal of hair in example (1) (i.e. *cl_HairRemoving)*, of tooth in example (2) (i.e. *cl_ToothRemoving),* and of device in example (3) (i.e. *cl_DeviceRemoving)*. The major problem here recurs in adjusting the degree of granularity of the model to the expressiveness of natural language. This is not a straightforward task as it is strongly dependent on the way such dictionaries will be used in potential applications. In particular, MLU applications have special requirements as described in the following section.

TUNING WORDS AND MEANINGS FOR MLU

Roughly speaking, the major task of MLU applications is to conciliate words with meanings.

Ambiguities and Domain Specific Language Uses

Ambiguities are mainly a problem for the analysis process, whereas the generation process is mostly concerned with domain specific language uses such as those adopted by health care professionals. One solution to lessen semantic ambiguities is to use domain-specific dictionaries that hold only meanings relevant to the treated domain. For example, the meaning of quarrel for the English word *rupture* will be discarded in the medical domain, whereas the interpretation as burst or as perforation will be emphasized respectively in the vascular surgery and oto-rhino-laryngology (ORL) domains. Another alternative, which allows a unique source for dictionaries to be maintained, is to add an additional argument to the dictionary specifying the domain of validity of the annotation. Besides, a simple solution to cope with usage is to define a preference order among all the possible annotations for a particular concept. In the current state of our dictionaries, the order selected to specify lexical entries defines an implicit preference. But this solution is not satisfactory when working with multiple dictionary sources as well as when no preferential rule can be set up a priori. For example, the order in which the two prefixes for *cl_Joint* are specified (see Figure 1), is consistent with the fact that *arthro* is commonly used to define compound words (e.g. *arthritis*, *arthrotomy*, *arthroscopy*), whereas *articulo* has a limited use (e.g. *articular surface*). This remark is valid for English and French. For expressing a *breast pain*, however, no formation rule can be set up between languages as this can be expressed in English by *mammalgia*, *mastalgia*, or *mastodynia* whereas only *mastodynie* is used in French. Finally, 'awkward' forms, i.e. semantically correct but never used in standard language, can be generated (e.g. **spondylopexy* for a *fixation of vertebrae*). A general solution to check the correctness of common compound words is to explicitly maintain a list of exceptions for each treated language.

This situation becomes more complex when ambiguities arise due to domain-specific language uses. Indeed, most of the surgical procedures, described in the CCAM classification, do not specify the adjective *surgical* on their labels as such information is clarified by the context. However, the words *excision* or *transplantation* must be interpreted as the concepts *cl_SurgicalExcising* or *cl_SurgicalTransplanting*, and not as *cl_Excising or cl_Transplanting*. Annotating *cl_SurgicalExcising* with the literal expression *surgical excision* is unsatisfactory from the common medical practice viewpoint. The annotation *excision* is equally unsatisfactory from the conceptual viewpoint, as it does not reflect the complete meaning of the concept. The solution chosen here was to handle usage directly in MLU tools and not in dictionaries. This implies specifying the kind of permitted implicit information (such as *cl_SurgicalRole*) and then inferring (especially in analysis) or masking (especially in generation) such information.

Other Intricacies of Natural Language

In order to go further into the quality of semantic dictionaries, let us consider a few intricate cases. For example, the stomatology rubric S291 speaks about a *temporo-mandibular arthrotomy* that is semantically represented by the concepts *cl_SurgicalIncising* and *cl_TemporoMandibularJoint* linked through the relationship *rel_actsSpecificallyOn*. The prefix *arthro* together with the adjective *temporo-mandibular* constitute a complete annotation for the concept *cl_TemporoMandibularJoint*. However, this is not manageable in our dictionary structure. The solution here is to exploit the compositionality of the concept *cl_TemporoMandibularJoint* by directly annotating its components (i.e. *cl_Joint*, *cl_TemporalBone* and *cl_Mandibule*). A close case occurs with the French expression *à visée diagnostique* (i.e. *with diagnostic aim*) where the preposition *with* plus the noun *aim* denote the relationship *rel_hasSpecificGoal* and the adjective *diagnostic* the concept *cl_DiagnosticAct*. Here, this excerpt of conceptual representation should be considered as the semantic representation for the full expression. Another case where medical jargon should be directly entered as an annotation of composite concepts, concerns the ORL rubric E678. Indeed, the French idiom *queue du sourcil* (i.e. *tail of eyebrow*) cannot be split into the concepts *cl_BodyRegion*, *cl_lateralSelection* and *cl_Eyebrow* effectively used to model its meaning as the lateral part of the eyebrow. Finally, there are contextual words for which it is difficult, a priori, to specify their full interpretation. This is the case for the adjective *parenchymatous* designating the set of functional tissues of an organ. In the neurology rubric N028 speaking about *a parenchymatous meningeal cutaneous lesion*, this adjective should be interpreted as *cl_Cerebrum* due to the presence of *cl_Meninges*. Such contextual cases are difficult to manage, as they require common sense knowledge.

RESULTS

The GALEN model, chosen as the semantic repository for our multilingual dictionaries, currently contains 13167 concepts. 5703 are composite concepts, among which 1568 remain anonymous. The automatic annotation process, set up to generate a 'lazy' English dictionary from the concept labels, has yielded 10515 lexical entries as it was adapted for dealing with theoretical concepts and usage considerations. Thereafter, our dictionaries have mainly grown following the experiments conducted with the University of Saint Etienne for the development and maintenance of the new French CCAM classification for surgical procedures. This classification is divided into 15 chapters, each containing about 500 French rubrics. In order to refine and finalize the linguistic labels of these rubrics, a generation in French, but also in English for comparisons, has been systematically performed from the GALEN representations set up to grasp the meaning of each rubric. Until now, 11 chapters have been entirely treated with a French dictionary of 3353 lexical entries of which 235 are affixes and an extra English dictionary of 806 lexical entries. The annotations were initially made by the domain experts at Saint Etienne, and are further refined during the linguistic review of generated sentences in Geneva. Finally, several small experiments have been carried out for testing our model-based approach with other European languages such as Italian, German, Spanish, Dutch, and Swedish.

CONCLUSION AND FUTURE WORK

This paper reports on the benefits and constraints of exploiting a concept model for building semantic dictionaries. It clearly highlights the fact that the tuning of linguistic and conceptual knowledge is a challenging process that should harmonize the breadth of uses with the depth of detail needed for the representation of medical data.

Although the quality of such annotations has proved to be adequate for use in MLU applications, the main bottleneck with such a model-based approach remains the size of the dictionaries, which are regulated by the model coverage. Indeed, the expansion of concept models, which is usually performed independently of any linguistic task, is often burdened with too much abstract modeling information. Alleviating such a modeling structure, especially at the level of the concept hierarchy, should lead to a more flexible system directly useful for large-scale MLU tasks. Such a strategy will soon be corroborated at the Geneva University Hospital, where a large project for automatic semantic indexing of electronic patient records is under way.

Acknowledgments

This work is funded by the Swiss government (OFES - Office Fédéral de l'Education et de la Science) as part of the European GALEN-IN-USE project.

References

1. Chute CG, Baud RH, Cimino JJ, Patel VL, Rector AL. Special Issue on Coding and Language Processing. Meth Inform Med, 1998; 37(4-5).
2. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR. The Content Coverage of Clinical Classifications. JAMIA, 1996; 3: 224-233.
3. Hirschman L, Sager N. Automatic Information Formatting of a Medical Sublanguage. In: Kittredge R and Lehrberger J (Eds.). Sublanguage: Studies of Language in Restricted Semantic Domains. Berlin: Walter de Gruyter, 1982: 27-80.
4. Narazenko A, Zweigenbaum P, Bouaud J, Habert B. Corpus-Based Identification and Refinement of Semantic Classes. In: Masys DR (Ed.). 1997 AMIA Annual Fall Symposium. Proceedings. Philadelphia: Hanley & Belfus, Inc., 1997: 585-589.
5. Baud R, Lovis C, Rassinoux A-M, Michel P-A, Scherrer J-R. Automatic Extraction of Linguistic Knowledge from an International Classification. In: Cesnik B, McCray AT, Scherrer J-R (Eds.). MEDINFO'98. Proceedings. Amsterdam: IOS Press, 1998: 581-585.
6. Zweigenbaum P, Courtois P. Acquisition of Lexical Resources from SNOMED for Medical Language Processing. In: Cesnik B, McCray AT, Scherrer J-R (Eds.). MEDINFO'98. Proceedings. Amsterdam: IOS Press, 1998:586-590.
7. Rector A. Compositional Models of Medical Concepts: Towards Re-usable Application-Independent Medical Terminologies. In: Barahona P and Christensen JP (Eds.). Knowledge and Decisions in Health Telematics. Amsterdam: IOS Press, 1994: 109-114.
8. Rassinoux A-M, Lovis C, Baud RH, Scherrer J-R. Versatility of a Multilingual and Bi-directional Approach for Medical Language Processing. In: Chute CG (Ed.). 1998 AMIA Annual Symposium. Proceedings. Philadelphia: Hanley & Belfus, Inc., 1998: 668-672.
9. Rodrigues J-M, Trombert-Paviot B, Baud R, Wagner J, Meusnier-Carriot F. Galen-In-Use: Using artificial intelligence terminology tools to improve the linguistic coherence of a national coding system for surgical procedures. In Cesnik B, McCray AT, Scherrer J-R (Eds.). MEDINFO'98. Proceedings. Amsterdam: IOS Press, 1998: 623-627.
10. Wolff S. The Use of Morphosemantic Regularities in the Medical Vocabulary for Automatic Lexical Coding. Meth Inform Med, 1984; 23(4): 195-203.