Chapter VI. GALEN BASED FORMAL REPRESENTATION OF ICD10

Abstract

Objectives: The main objective is to create a knowledge-intensive coding support tool for the International Classification of Diseases (ICD10), which is based on formal representation of ICD10 categories. Beyond this task the resulting ontology could be reused in various ways. Decidability is an important issue for computer assisted coding, consequently the ontology should be represented in description logic.

Methods: The meaning of the ICD10 categories is represented using the GALEN Core Reference Model. Due to the deficiencies of its representation language (GRAIL) the ontology is transformed to the quasi-standard OWL. A test system which extracts disease concepts and classifies them to ICD10 categories has been implemented in Prolog to verify the feasibility of the approach.

Results: The formal representation of the first two chapters of ICD10 (infectious diseases and neoplasms) has been almost completed. The constructed ontology has been converted to OWL DL. The test system successfully identified diseases in medical records from gastrointestinal oncology (84% recall, however precision is only 45%). The classifier module is still under development. Due to the experiences gained during the modelling, in the future work FMA is going to be used as anatomical reference ontology.

1. Introduction

Indexing of medical diagnoses is a difficult and error-prone task. Providing assistance to manual coding is an important research area in medical informatics since many decades [1], and despite all efforts the problem has not yet been solved. Computer assisted coding systems can be basically classified into two groups:

- **Systems using statistical methods** do not “know” anything about the coding systems and the natural language, they classify the diagnoses based on statistical features of learning samples [2-4]. Such systems are language-independent and easy to implement: only a well-controlled training sample in the given language is required. The usage of thesauri (replacing synonyms with preferred term, etc.) can significantly enhance the performance [5]. In case of some languages (e.g. German, Hungarian) stemming also increases performance. The drawback of statistical methods is that their effectiveness strongly depends on the size and completeness of the training sample.

- **Systems using knowledge-intensive methods** represent both the coding system and the clinical text written in natural language. The creation of the knowledge base is a resource intensive task, but – at least
in principle – the formal representation of medical narratives can support the reuse of information in various ways (clinical decision support, exchange of information among different EHR systems, etc.) When the knowledge base describes both the clinical concepts and the categories of the coding scheme, the system can infer the appropriate codes even in cases where the clinical expression uses different terms or even different concepts than the coding scheme. At the moment, the most popular knowledge representation is based on the use of ontologies.

This paper presents a knowledge-intensive method for assisting ICD10 (International Statistical Classification of Diseases and Related Health Problems, revision 10) [6] coding. First the most important features of ICD10 are presented, then we describe the GALEN Core Reference Model (CRM) briefly. After that the formal representation of ICD10 is presented, with examples highlighting some peculiarities. Finally the test system implemented in Prolog is described, followed by a summary of the achieved results.

2. Material and methods

2.1. ICD10 classification system

ICD10 is the most frequently used classification of diseases in Europe. It has been published in 1992 by WHO in three volumes. Our work is based on the first volume, which contains the ICD codes together with their natural language labels, definitions, (local) coding rules, etc. The second volume defines global coding rules and the third volume is a mere index to the first volume.

ICD10 is a hierarchical coding system, organised into 5 levels:
• The 21 chapters group together diseases according to major categories (location – e.g. cardiovascular diseases and pathology – e.g. neoplasms). A separate chapter contains the international classification of oncology based on SNOMED [7].
• Chapters contain sections grouping together similar diseases (like J10-J18 “Influenza and pneumonia”).
• Sections contain groups, which collect very similar diseases (like J10 “Influenza due to identified influenza virus”). Groups are defined by three character codes.
• Groups contain items, which define narrow groups of diseases (like J10.0 “Influenza with pneumonia, influenza virus identified”). Items are defined by four character codes.
• In some cases items are subdivided on the fifth character. This subdivision is left for national purposes (e.g. H4411 “Endogenous uveitis” and H4419 “Other endophthalmitis nos.”), however WHO itself defined also some
categories. For example at S02 “Fracture of skull and facial bones“:

The following subdivisions are provided for optional use in a supplementary character position where it is not possible or not desired to use multiple coding to identify fracture and open wound; a fracture not indicated as closed or open should be classified as closed:

0 closed
1 open.

Each category has a name and may have (local) coding rules, definitions, and comments. Groups and items may have synonyms and dagger-asterisk cross-references. Since our goal is to formally represent the meaning of the given category, these pieces of information are not represented explicitly in the ontology. However, they have to be taken into account during modelling, since they may affect the meaning of the category.

Since the aim of ICD10 is to classify all possible diseases, there are lot of “other” (e.g. J10.8 “Influenza with other manifestations, influenza virus identified”) and “not otherwise specified” (e.g. J12.9 “Viral pneumonia, unspecified”) categories in the coding system. The latter does not cause any problems (since it can be modelled as “Viral pneumonia”), however the formal representation of “other” categories is a difficult task. The problem is that the meaning of “other” is not necessarily equivalent with the intersection of the parent concept with the complement of the explicit sibling categories. Other” (or “not elsewhere classified”) means something that cannot be found elsewhere in the whole classification. For that reason a detailed review of the coding system is required to understand what actually is meant by it. The underlying problem is that ICD is a compromise on various possible aspects of classification (e.g. cause versus localisation of the disease), consequently various forms of a disease may appear in totally different parts of the classification. For example, diabetes mellitus is located in the chapter of endocrine diseases, but diabetes associated with pregnancy appears in the chapter of complications of pregnancy. This phenomenon is one cause why clinicians have difficulties using ICD10.

2.2. **GALEN CRM**

The GALEN Core Reference Model (CRM) is designed to be a reusable, application- and language-independent model of medical concepts to support EHRs, clinical user interfaces, decision support systems, classification and coding systems, etc. [8]. The development of GALEN was started about 10 years ago and the computational resources of that time limited the semantics of its description language, GRAIL (GALEN Representation and Integration Language).

The GALEN CRM is a ‘third generation’ terminology system [9]. The key feature of the GALEN approach is that it provides a model – a set of building blocks and constraints – from which concepts can be composed (in contrast to
traditional classification systems). The classification of composite concepts is automatic and based on formal logical criteria.

The GALEN CRM has four logical layers: high-level ontology, CRM, subspeciality extensions and model of surgical procedures. For the representation of ICD10 we need only the first three layers.

GRAIL is a description logic-like language [10] with special features to handle part-whole relations and other transitive relations needed to represent medical knowledge. In description logic (DL) these features are called role propagation and role composition and the existing DL reasoning systems began to support them only recently [11]. GRAIL shows some similarity also to conceptual graphs [12], and to typed feature structures [13]. However, GRAIL has evolved separately since late 1980s, driven by clinical applications. Even so, GRAIL is typically referred to in the literature as a description logic language. The main differences between GRAIL and a typical DL language are:

- GRAIL uses a so called grammatical/sensible level sanctioning, which is similar to the idea of canonical graphs in conceptual graph theory. This notion means that an attribute (a role in DL) can only be used after it has been allowed (contrary to the “open world assumption” in DLs where any relation can be used unless the domain and range constraints or “local” property restrictions prohibit it). A grammatical sanction is a statement that “an abstraction is useful for querying but not sufficiently constrained for generation”. A sensible sanction means that the constraint can be used for generating complex concepts, but first it has to be permitted on the grammatical level. In DL formalism the checking of whether a particular sensible sanctioning is valid would require the induction of additional relations.

- GRAIL has no concept constructors without quantification. It is possible to represent sets (by the concept collection) and there is a workaround for limited negation (introduced to represent conditions with or without other conditions, but it has been implemented in the ontology, not in the grammar).

GRAIL has a construct called refinedAlong (or specialisedBy). In fact, it is role propagation, which is a feature of some DL languages. The semantics of role propagation is:

\[ r \circ s = \forall x, y, z: r(x, y) \text{ AND } s(y, z) \rightarrow r(x, z) \]

where \( r \) and \( s \) are the relations, and \( x, y \) and \( z \) are classes.

\begin{align*}
\text{hasLocation} \circ \text{partOf} & \subseteq \text{hasLocation}
\end{align*}

which allows the reasoner to infer e.g. that a heart valve disease is a heart disease (since heart valve is a part of the heart). This is a very important peculiarity of medical knowledge representation. Moreover it can be used as a tool connecting the detailed description of a diagnosis to a more abstract disease category in ICD.
Chapter VI – GALEN based formal representation of ICD10

2.3. GALEN based formal representation of ICD10

The goal is to represent formally the meaning of each ICD category, using the concepts and the attributes of GALEN CRM. The ontology of ICD10 contains only concepts representing each ICD10 category (named by the ICD code), and the formal definition of the category. Other information (coding rules, references, etc.) belonging to a certain ICD category provides only additional information to the user of the coding system. Therefore it is advisable to separate it from the formal representation. The hierarchical relations of ICD10 are neither represented in the ontology, since they not necessarily always overlap with the hierarchy inferred from the formal definitions.

According to our view the ICD categories can be defined by a multi-axial classification. The main axes are:

- Anatomy: location of the disease (if applicable).
- Morphology: type of lesion.
- Etiology: cause of the disease (if applicable).

This view is similar to that of SNOMED version 2 and 3, which is the pathologist’s view of medicine.

The anatomy axis contains the anatomical entities of the human body (in case of ICD these are tissue, organ parts, organs, organ systems and regions). The morphology axis contains the types of pathological lesions, such as inflammations and neoplasms. The etiology axis contains the causative agents of the disease: organisms, chemical, physical and socio-environmental entities.

In case of certain categories additional axes are required, such as mode of diagnoses (e.g. A15 “Respiratory tuberculosis, bacteriologically and histologically confirmed”). We have to remark that such a distinction has nothing to do with the ontology of tuberculosis, since it is a diagnostical aspect, and in that sense it is outside of the original aim of ICD: to classify diseases.

The mode of disease transmission is modelled in the organism axis. For example, A83 “Mosquito-borne viral encephalitis” is represented as Encephalitis which isSpecificConsequenceof (ArboVirus which isActedOnBy (Transmitting which hasSpecificPersonPerforming Mosquito)). The disease may have complications (A98.5 “Hemorrhagic fever with renal syndrome”) or be itself a complication of another disease (B01.0 “Varicella meningitis”).

These peculiarities require that the category has to be defined by formal relations among potentially complex entities. Our decision was to define a category by the following template:
(MorphologicalEntity which <
    hasLocation AnatomicalEntity
    isConsequenceOf (EtiologicalEntityOrDisease)
    isIdentifiedBy DiagnosticProcedure
    hasComplication Disease
    modifierAttribute ModifierConcept >) name
DiseaseCategoryCode.

In some cases GALEN contains a composite entity of anatomy and
morphology (e.g. meningitis). In this case the composite entity is used. A relation
may be omitted if there is no constraint on that particular relation. The entities
may be complex classes not present in the underlying core ontology (e.g.
“ArboVirus which isActedOnBy (Transmitting which hasSpecificPersonPerforming Mosquito)). The required modalities of the diseases
(such as chronicity, laterality, disease state, mode of acquisition) are defined by
using GALEN modifier concepts.

For example, A81.1 “Subacute sclerosing panencephalitis”, which is an
autoimmune disease caused by measles infection is represented as:

Encephalitis which isConsequenceOf
    (AutoimmuneProcess which isConsequenceOf (InfectionProcess
        which isSpecificConsequenceOf MeaslesVirus))

2.4. Transforming the ontology into OWL

The “not otherwise classified” categories can be represented as the intersection
of the parent concept with the complement of the logical sibling categories. The
parent concept is not always a defined ICD10 category (e.g. in case of A02 “Other
salmonella infections”, since there is no “Salmonella infections in general” in
ICD10). Intersections and complements cannot be represented in GRAIL.

Another limitation of the practical use of GRAIL is the lack of publicly available
reasoner.

These two reasons led us to convert the ontology to the quasi-standard Web
Ontology Language (OWL) [14]. We found that OWL DL (which belongs to the
DL language class SHOIN) is the best practical choice for our task. The only
important limitation is the lack of support of role propagation.

Since there are cases where the reasoning on part-whole relation is important
for ICD10 coding (e.g. clinical diagnosis “Fracture of tibial head” to be mapped to
S82.1 “Fracture of upper end of tibia”) the role propagation axioms have to be
added to the ontology. It is advisable to store them in the OWL file as
AnnotationProperty-s. Therefore the resulting OWL file is valid, available OWL
reasoners can load it, and can reason about it (except role propagation based
reasoning). If an own special OWL interface is implemented above the reasoner, the system can also take the role propagation axioms into account.

Another solution is to create a work-around on role propagation [15]. This approach adds the necessarily “rules” not at the definition of roles, but at the concept level, which makes it possible to use common reasoners. Moreover if role propagation is defined for a pair of relations then it holds for every occurrence. (e.g. if there is role propagation for hasLocation and hasPart, than Appendicitis will be a kind of Enteritis, since appendix is a part of intestine. This is not what physicians would think.) By moving the assertion to the level of concept definitions the ontology engineer can decide whether such a subsumption is true, and can add the necessary relations. However (s)he has to review each case, which increases the time to create the ontology.

For example, the category “fracture of femur” would be defined as a fracture located at the femur. However since the “fracture of a part of femur” is a “fracture of femur” the category should be rather defined as a fracture which is located at the femur, or at some part of it.

The following GRAIL constructs have been transformed to OWL:

- The newSub/addSub/addSuper operators are used to define asserted subsumption. Their corresponding element in OWL is rdfs:subclassOf.
- The operators which/whichG formally define a category. We made no distinction between them during the transformation, since whichG defines complex entities used for querying. The “A which hasX B” concept is represented in OWL by the intersection of A with the ObjectProperty hasX restricted to B.
- The necessarily/topicNecessarily/valueNecessarily operators express the necessity of the criterion. valueNecessarily is the inverse of topicNecessarily and necessarily asserts both criteria. A topicNecessarily B is converted to a class (A) which is a subclass of an unnamed class with property restriction owl:someValuesFrom.
- The specialisedBy (refinedAlong) construct has been transformed as an AnnotationProperty. The “AttrA refinedAlong AttrB” assertion is represented as a DatatypeProperty annotation containing the string “AttrB” added to AttrA (since ObjectProperty must refer to objects, not to properties).

Since sanctioning is a tool supporting only concise modelling it has not been converted to OWL. It is possible to convert sensible level sanctioning to owl:allValuesFrom. The relation between grammatical and sensible level sanctioning cannot be represented in DL.

The definition of “other” categories is achieved by the owl:disjoint construct. The category is a subclass of the parent category and the owl:disjoint relation is asserted between the “other” category, and each of its (logical) siblings. (The other solution – with the same meaning – would be the use of owl:complementOf in this case the category would be the intersection of the parent category with the complement of the union of its siblings.)
2.5. **Automatic coding tool**

The natural language processing (NLP) module of the system is a simple statistical component augmented with a thesaurus, since we only need to identify the expressions denoting diseases in texts. To allow an easy implementation, the domain of the medical records has been restricted to gastrointestinal oncology. The whole textual content of the medical records, not only the “clinical diagnoses” field is searched for disease names. The sentences are analysed almost independently, no anaphora resolution is performed, only terms of a disease name distributed over adjacent sentences are merged. The module contains a dictionary providing translation of Hungarian and Latin names of anatomical and morphological terms of the domain into English (the labels of GALEN concepts are typically in English).

First the text is broken down to sentences with boundaries identified by the sequence “period-space-capital letter”, then the words are translated with the dictionary. Morphological analysis is not performed, only statistical similarity of the given word compared to the words in the thesaurus is computed using a modified version of Levenshtein distance. Candidates are ranked, and only the most relevant ones are considered. The found anatomical and morphological entities are combined into a notion of disease, which is described in GRAIL. These diseases are displayed, together with the originating sentences and relevant words.

The medical records were manually analysed, the relevant diseases have been extracted, and manually coded to ICD10, allowing us to verify the performance of the NLP module. The module identifies 84% of the relevant diseases, however it also found a lot of unnecessary diseases and locations, consequently the precision is low: 45%.

The coding module – which classifies the found disease concepts – is still under development. The idea is that the found concepts are classified by the SILK DL reasoner [16] into ICD10 categories. If the two concepts are not totally identical, a similarity measure can be estimated. In most cases the found disease concept is a subclass of one or more ICD10 categories. However this subsumption relation can frequently only be identified by using role propagation. Since the used DL reasoner cannot cope with role propagation, we are currently redefining the ICD categories according to the work-around method described in Section 2.4. Another solution would be to augment the reasoner with this feature, however the work-around leaves more freedom to the ontology engineer, and we think it should be used even if the creation of the ontology is more time consuming.
3. **Discussion**

The formal definition of the first two chapters of ICD10 (infectious diseases and neoplasms) has been almost completed. During the creation of the ontology only the hierarchy of ICD categories has been taken into account for detecting sibling concepts. After the completion of the formal representation of the whole ICD10 a review step will be required to find parent categories (subsumption relations) that are not represented explicitly in the ICD hierarchy. A formal consistency check is also necessary since ICD is not supposed to be consistent.

The work-around on role propagation seems promising, therefore we are currently redefining the ICD categories according to it. This is a time consuming task, since the ontology engineer has to analyse each ICD category, to find out whether it is sensible to speak about the parts of the particular anatomical structure, and whether the given morphological lesion located at any of these parts is a subtype of the given category. This task can be supported by a software system displaying the part-whole taxonomy and the ICD categories related to the concepts in this taxonomy.

During the work some problems have been found with GALEN CRM, as core ontology. First, some of the required anatomical (“retroperitoneal lymphnode”) and many etiological (“enteropathogenic Escherichia Coli”) concepts were missing from GALEN CRM. These concepts were added to the core ontology. Second, there were some concepts that could not be defined using the concepts and the attributes of GALEN CRM.

- The meaning of C10.4 “(malignant neoplasm of) branchial cleft” is a malignant neoplasm located in a branchiogen cyst, which is a developmental residuum. The definition of such concepts in GALEN would require the construction of (human) developmental anatomy, with the required concepts for temporary phenomena. Moreover a branchiogen cyst persisting in postnatal life is in itself a pathological structure.
- The representation of C06.2 “(malignant neoplasm of) retromolar area” requires the definition of “retromolar region”. For that, new attributes describing 3D relations would be needed.

Due to these problems, it has been decided that the project will be continued using the Foundational Model of Anatomy (FMA) as core anatomy ontology [17]. FMA is a detailed ontology of anatomy and human development. The work up to now would not be lost; most of the references to anatomical concepts in GALEN can be automatically converted to references to concepts in FMA.

The conversion of FMA from the Protégé-2000 RDF format to OWL DL is underway. The modification of the conceptual system is also necessary:

- To facilitate reasoning the ontology should be as compact as possible, therefore the enumerative structures in FMA (e.g. “Radiate ligament of head of first rib”) have to be transformed to a composite approach (such
as that of GALEN). We hope that with this transformation the number of concepts could be reduced from approximately 70,000 to about 10,000 without losing the information contained in FMA.

- The ontology should be aligned to a formal top-level ontology, DOLCE Lite [18] has been chosen for this task. We have chosen this ontology because it is the best top-level ontology written in OWL DL known by us. The use of a formal top-level helps to maintain the consistency of the ontology, and makes reasoning feasible.
- The conceptual systems describing physiology, pathology, etc. also have to be created under the same top-level.
- Some problems with ICD10 (in an ontological sense) have also been found:
  - C09 “Malignant neoplasm of tonsil”, and C09.0 “Tonsillar fossa”. The tonsillar fossa is not a part of the tonsil, it is normally occupied by the tonsil. A neoplasm may occur in it only after removal of the tonsil. It is not a serious problem; it only shows that the hierarchical relations in ICD not necessarily coincide with formal subsumption.
  - In case of C76-C80 “Malignant neoplasms of ill-defined, secondary and unspecified sites” the formal definition of “ill defined site” is not possible (however its children can be defined.). It is again a diagnostic problem: the concrete location of the neoplasm is not known, only the region in which it is located.

The extension of the system to other neoplastic diseases is under way, however first the thesauri have to be extended to achieve adequate performance.

4. Conclusion

It has been proven that it is technically possible –with some rare exceptions – to represent ICD categories formally using GALEN CRM converted to OWL DL.

A relatively simple NLP tool is able to effectively extract disease names from medical narratives and to represent them formally. Such a simple tool works with acceptable recall and low precision, consequently human control is not avoidable.

It seems to be advantageous to use FMA as a reference model of anatomy, however it has to be compacted and harmonised to a formal top-level ontology for the formal representation of other medical entities.

Acknowledgements

This work has been partially founded by the Hungarian Ministry of Education, contract No. IKTA 00126/2002 and by the Hungarian Ministry of Health. The kind help of dr. Gábor Csongor Kovács is highly appreciated.
References


[8] Information materials about Galen can be found at http://www.opengalen.org/open/crm/index.html


[14] Information materials about OWL can be found at http://www.w3.org/2004/owl


[17] Information material about FMA can be found at http://sigpubs.biosstr.washington.edu/view/projects/Foundational_Model_of_Anatomy.html
[18] Information material about DOLCE can be found at: http://www.loa-cnr.it/DOLCE.html