Connecting Medical Informatics and Bio-Informatics
R. Engelbrecht et al. (Eds.)
ENMI, 2005

GALEN Based Formal Representation of ICD10

Gergely Héja[a], György Surján[b], Gergely Lukácsy[c], Péter Pallinger[a], Miklós Gergely[a]

[a] *Budapest University of Technology and Economics, Department of Measurement and Information Systems*
[b] *National Institute for Strategic Health Research*
[c] *Budapest University of Technology and Economics, Department of Computer Science and Information Theory*

Abstract

The authors present a formal representation of ICD10 based on GALEN CRM. The goal of the work is to create a coding support tool for coding clinical diagnoses to ICD10. The formal representation of the first two chapters of ICD10 has been almost completed. The paper presents the main aspects of the modelling, and the experienced problems. The constructed ontology has been converted to OWL, and a test system has been implemented in Prolog to verify the feasibility of the approach. The system successfully identified diseases in medical records from gastrointestinal oncology. The classifier module is still under development.

Keywords:
Ontology, ICD10, GALEN

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], still unsolved. Computer-assisted coding system can be basically classified into two groups.

Statistical systems do not "know" anything about the coding systems and the natural language, they classify the diagnoses based on statistical features of the training samples [2, 3]. Such systems are language-independent and easy to implement, since only well-controlled training samples are required. The usage of thesauri could significantly enhance the performance of such systems [4]. The drawback of this approach is that it can only cope with problems more or less masked by the training sample.

Knowledge-intensive systems represent formally both the coding system and the clinical text to be coded. The creation of the knowledge base is a resource intensive task, but the knowledge-based formal representation of medical narratives can support the reuse of information in various ways (clinical decision support, communication between different EPR systems, etc.) When the knowledge base describes both the clinical concepts and those of the coding system, the system can infer the possible codes even in those cases when the clinical expression uses different terms or even different concepts than the code category.

This paper presents a knowledge-intensive method for assisting ICD10 [5] coding. Both manual and computer-assisted coding processes may use clinical diagnoses as input information. This is a rational constraint (although it has some drawbacks [6]), because processing of the whole patient record would require a very complex model. In cases where the diagnosis is not specific enough, the user should consult the patient record.

2. Material and methods

2.1. ICD10

ICD10 is the most frequently used classification of diseases in Europe. It has been published in 1992 by WHO in 3 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 in 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 contain items, which define narrow groups of diseases (like J10.0 "Influenza with pneumonia, influenza virus identified"). In some cases items are subdivided on the fifth character (like H4411 "Endogenous uveitis"). 5th character subdivision is left for national purposes, however WHO itself defined some categories.

Each category has a name and may have (local) coding rules, definitions, and comments. Groups and items may have synonyms and dagger-asterisk cross-reference. Since our goal is to formally represent the meaning of the given category, these pieces of information is not directly represented in the ontology, but 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 a 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. The later does not cause any problems, however the formal representation of "other" is a more difficult task. The problem is that it is not certain that an "other" category is defined by the exclusion of its siblings from its explicit parent category, therefore a detailed review of the coding system is required. The cause of this phenomenon is that the localisation of disease groups with similar (clinical) meanings may be in completely different parts of the hierarchy.

2.2. GALEN

The GALEN Core Reference Model (CRM) is designed to be a reusable application-independent and language-independent model of medical concepts to support EHRs, clinical user interfaces, decision support systems, classification and coding systems, etc [8]. 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 structure of the GALEN has four logical layers: high-level ontology, CRM, subspecialty extensions and model of surgical procedures. For the representation of ICD10 only concepts from the first three layers were needed.

GRAIL is a description logic-like language [9] with special features to handle part-whole relations and other transitive relations needed to represent medical knowledge. These features are called role propagation and role composition in the description logic world, and the existing DL reasoning systems began to support them only recently [10]. It is related also to Conceptual Graphs, and to typed feature structures. 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 multi-level sanctioning and the lack of concept constructors without quantification.

Sanctioning is similar to canonical graphs in conceptual graph theory [11]. This notion means that an attribute (a role in DL) can only be used after it is allowed (in contrary to DLs where any relation can be used if it is not prohibited). Grammatical sanction is a statement that "an abstraction is useful for querying but not sufficiently constrained for generation". Sensible sanction means that the constraint can be used for generating complex concepts, but first it has to be permitted on grammatical level.

The lack of concept constructors without quantification is partially solved by elements in the ontology. Sets are represented by the concept **Collection** and there is a workaround for limited negation (introduced to represent conditions *with* or *without* other conditions, but it is implemented in the ontology, not in the grammar).

The role propagation in GRAIL can be asserted by the construct called *refinedAlong* (or *specialisedBy*), with the semantics:

$$r \circ s \equiv \forall\ x, y, z: r(x, y) \text{ AND } s(y, z) \rightarrow r(x, z)$$

where *r* and *s* are the relations, and *x*, *y* and *z* are classes. If this axiom is asserted between *hasLocation* and *partOf*, the reasoner can infer e.g. that a heart valve disease is a heart disease. This is a very important peculiarity of medical knowledge representation.

2.3. GALEN based formal representation of ICD10

The goal is to represent formally the meaning of each ICD category, using the concepts and attributes of GALEN CRM. Only the formal definition of the category (labelled by the ICD code) is represented. Since the other information (e.g. coding rules) give only additional information to the user, therefore they are separated from the formal representation. The hierarchical relations of ICD10 are also not represented in the ontology, since they may not always overlap with the hierarchy inferred from the formal definition.

According our view ICD categories can be defined by a multi-axial conceptual system:

- anatomy: location of the disease (if applicable). In case of the ICD anatomical entities are tissues, organ parts, organs, organ systems and regions.
- morphology: type of pathological alteration (e.g. inflammations and neoplasms)
- etiology: cause of the disease (if applicable, mostly organisms, chemical, physical and socio-environmental entities)

This classification is similar to SNOMED, which is the pathologist's view of medicine.

In case of certain categories additional axes are required, such as mode of diagnosis (e.g. A15 "Respiratory tuberculosis, bacteriologically and histologically confirmed"). By the way, this additional constraint has nothing to do with what is tuberculosis, thus it could be seen as a violation to original aim of ICD: to classify diseases.

Organisms can be further specified by e.g. type of transmission (A83 "Mosquito-borne viral encephalitis"). The disease may have complications (A98.5 "Hemorrhagic fever with renal syndrome") or be itself a complication of another disease (B01.0 "Varicella meningitis").

Thus categories are defined by formal relations among potentially complex entities:

MorphologicalEntity which
hasLocation **AnatomicalEntity**
isConsequenceOf **EtiologicalEntityOrDisease**
isIdentifiedBy **DiagnosticProcedure**
hasComplication **Disease**
modifierAttribute **ModifierConcept**

In some cases GALEN contains a composite entity of anatomy and morphology (e.g. meningitis). A relation may be omitted if there is no constraint on the 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, acquisition mode, etc.) could be defined by using GALEN modifier concepts.

E.g. A81.1 "Subacute sclerosing panencephalitis", which is an autoimmune disease caused by measles infection is represented as:

Encephaliti*s* which *isConsequenceOf*
(**AutoimmuneProcess** which *isConsequenceOf* (**InfectionProcess** which *isSpecificConsequenceOf* **MeaslesVirus**))

2.4. Transforming the ontology to OWL

The aforementioned "other" categories can be defined by the exclusion of its (logical) siblings from the parent concept. 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 ICD10). This construction of exclusion cannot be represented in GRAIL.

This reason and the problems with reasoning support for GALEN led us to convert the ontology to the quasi-standard OWL [12]. We found that OWL is appropriate for the task, except for the definition of role propagation. Nevertheless we have chosen OWL, because we expect that role propagation would be added to it shortly.

Since there are cases when the reasoning on part-whole relation is important for ICD10 coding (e.g. S62.1 "Fracture of other carpal bone(s)" – clinical diagnosis "Fractura ossis lunati") the role propagation axioms have to be added to the ontology. It is advisable to store them in the OWL file, in the ontology header. 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 OWL interface is implemented above the reasoner (which supports role propagation), the system can also take the role propagation axioms into account.

The following GRAIL constructs have been transformed to OWL:

- The newSub / addSub / addSuper operators are used to define asserted subsumption. Their parallel in OWL is rdfs:subclassOf.
- The operators which / whichG formally define a category. We made no distinction between them in OWL. The "**A** which *hasX* **B**" concept is represented in OWL by the intersection of **A** with the (to **B**) restricted ObjectProperty *hasX*.
- The necessarily / topicNecessarily / valueNecessarily operators express the necessity of the criterion. ValueNecessarily is the inverse of topicNecessarily and necessarily asserts both criteria. **A** topicNecessarily *hasX* **B** is converted to a class (**A**) which is a subclass of an unnamed class with property restriction owl:someValuesFrom on the ObjectProperty *hasX*.
- Since sanctioning is only a tool supporting concise modelling it has not been converted to OWL yet. It is possible to convert sensible level sanctioning to owl:allValuesFrom.

The definition of "other" categories is achieved by the owl:disjoint construct.

2.5 Automatic coding tool

The NLP module of the system is a simple statistical component augmented with a thesaurus, since it has only to identify the expressions denoting diseases in texts. To allow easy implementation, the domain of the medical records has been constrained to

gastrointestinal oncology. The whole text is searched for disease names, not only the "diagnoses" field. The sentences are analysed almost separately, no anaphora resolution is performed, only terms of a disease name located in adjacent sentences are contracted. The module contains a dictionary that translates Hungarian and Latin names of anatomical and morphological terms of the domain into English (which is the typical language of labels in GALEN). The relevant anatomical and morphological entities are stored in two lists.

First the text is broken down to sentences (boundary identified by the sequence "period-space-capital letter"), then the words are translated by the dictionary. Morphological analysis is not performed, only the statistical similarity of the word compared to the words in the thesaurus is computed. The candidates are ranked, and only the most relevant ones are considered. The found anatomical and morphological entities are combined into a disease, which is described in GRAIL. The 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, thus allowing us to check the abilities of the NLP module. The module identifies 84% of the relevant diseases, however it also founds a lot of unnecessary diseases and locations, thus the precision is low: 45%.

The coding module – which classifies the found disease concepts into ICD10 categories – is still under development. The idea is that the found concepts are classified by the SILK DL reasoner [13] 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, frequently this subsumption relation can only be found out using role propagation. The used DL reasoner cannot cope with role propagation; therefore first of all it has to be augmented with this feature.

3. Results and discussion

The formal definition of the first two chapters of ICD10 (infectious diseases and neoplasms) has been almost completed. During the building of the ontology only the hierarchical relations have been taken into account as sibling concepts. After the completion of the formal representation of the whole ICD10 a review step is required to find the other related categories and the consistency errors.

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

- The meaning of C10.4 "(malignant neoplasm of) branchial cleft" is a malignant neoplasm located on branchiogen cyst, which is a developmental residuum (thus a pathological structure). The definition of such concepts in GALEN would require the defoinition of (human) development, with concepts for temporary phenomena.
- The representation of C06.2 "(malignant neoplasm of) retromolar area" requires the definition of "retromolar region", which needs attributes describing 3D relations.

Based on these problems, we have decided that the project will be continued using FMA as core anatomy ontology [14]. FMA is a detailed ontology of human anatomy and development. The work up to now would not be lost: most of the references to anatomical concepts can be converted automatically. The conversion of FMA to OWL DL is underway. For computational effectiveness the enumerative approach of FMA is transformed to a composite approach (such as that of GALEN). The modeling of physiology, pathology, etc. is also required.

Some problems (in ontological sense) with ICD10 have also been found:

- C09 “Malignant neoplasm of tonsil”, and C09.0 “Tonsillar fossa”, which is not a part of the tonsil, but a structure formed by tonsillectomy. This example shows that the hierarchical relations in ICD not necessarily coincide with formal subsumption.
- In case of C86-C90 “Malignant neoplasms of ill-defined, secondary and unspecified sites” the formal definition of “ill defined site” is not realisable.

4. Conclusion

The formal representation of ICD10, together with the required NLP and inference tools could significantly enhance the quality of coding. The first step to this aim has been fulfilled: the development of a model of ICD10 based on GALEN. The formal definition of two chapters of ICD10 has been almost completed, with observations indicating that GALEN may not be the appropriate core ontology. A test system has been created to automatically classify oncological clinical diagnoses. The statistical NLP method has good recall, however the precision is quite low. The classifier module is still under development. The results indicate that the goal is realisable, although the used inference engine has to be supplemented by role propagation, and a more efficient NLP module should be used.

5. 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.

6. References

[1] F. Wingert An indexing system for SNOMED, Meth Infom Med 1986; 25:22-30
[2] C. G. Chute, Y. Yang An overwiew of Statistical Methods for the Classification and Retrieval of Patient Events Methods of Information in Medicine (1995); 34:104-110
[3] G.Surján, G. Héja Maintenance of self-consistency of coding tables by statistical analysis of word co-occurrences Stud Health Technol Inform. 1999; 68:887-90.
[4] G. Héja, G. Surján: Analysing coding tables with thesaurus augmented statistical methods in CD ROM of MIE2003, XVIIIth International Congress of the European Federation for Medical Informatics, St. Malo, France, 4-7 May 2003.
[5] International Statistical Classification of Diseases and Health Related Problems, Who, Geneva, 1992
[6] G. Surján: Questions on validity of International Classification of Diseases-coded diagnoses, International Journal of Medical Informatics 54 (1999) 77-95
[7] Coté R.A. Rothwell D.J. et al. (eds.) SNOMED International, College of American Pathologists, Northfield Il. USA 1993
[8] Information materials about Galen can be found at http://www.opengalen.org/open/crm/index.html
[9] F. Bader et al. (editors): The Description Logic Handbook (Theory, Implementation and Applications), Cambridge University Press, 2003
[10] I. Horrocks, U. Sattler: Decidability of SHIQ with complex role inclusion axioms. Artificial Intelligence, 160(1-2):79-104, December 2004.
[11] J. F. Sowa Conceptual Structures: Knowledge Representation in Mind and Machine, John Wiley & Sons, New York, 1985
[12] Information materials about OWL can be found at http://www.w3.org/2004/owl
[13] T. Benkő, G. Lukácsy, A. Fokt, P. Szeredi, I. Kilián, P. Krauth: Information Integration through Reasoning on Meta-data, Proceedings of AI Moves to IA, Workshop on Artificial Intelligence, Information Access, and Mobile Computing", IJCAI'03, Acapulco, Mexico
[14] Information material about FMA can be found at http://sigpubs.biostr.washington.edu/view/projects/Foundational_Model_of_Anatomy.html

Address for correspondence

Budapest University of Technology and Economics, Dept. of Measurement and Information Systems,
2. Magyar tudósok körútja, Budapest, Hungary, H-1117, e-mail: heja@mit.bme.hu