Using Non-Primitive Concept Definitions for Improving DL-based Knowledge Bases

Ronald Cornet, Ameen Abu-Hanna
Department of Medical Informatics
Academic Medical Centre, Amsterdam, The Netherlands
{r.cornet,a.abu-hanna}@amc.uva.nl

April 29, 2004

Abstract

Medical Terminological Knowledge Bases contain a large number of primitive concept definitions. This is due to the large number of natural kinds that are represented, and due to the limits of expressiveness of the Description Logic used. The utility of classification is reduced by these primitive definitions, hindering the knowledge modeling process. To better exploit the classification utility, we devise a method in which definitions are assumed to be non-primitive in the modeling process. This method aims at the detection of: duplicate concept definitions, underspecification, and actual limits of a DL-based representation. This provides the following advantages: duplicate definitions can be found, the limits of expressiveness of the logic can be made more clearly, and tacit knowledge is identified which can be expressed by defining additional concept properties. Two case studies demonstrate the feasibility of this approach.

1 Introduction

Medical terminological knowledge bases (TKBs) represent knowledge about concepts, relationships and terms, in the domain of medicine. For example, a concept may be defined as "inflammation of the membranes of the brain or spinal cord", and described by the synonymous terms "cerebrospinal meningitis" and "meningitis". TKBs provide an invaluable source of structured medical knowledge, serving a range of purposes.

Advantages of representing this knowledge using Description Logics (DL) include the explicit semantics of the represented knowledge and the possibility to perform automatic reasoning based on this knowledge. The prominent reasoning tasks are satisfiability and subsumption. To infer subsumption, concept definitions with necessary and sufficient conditions are required. We will refer to such definitions as non-primitive definitions (sometimes referred to by others as "equalities"), whereas a primitive concept definition (sometimes referred to by others as "inclusion") specifies only necessary conditions. It is however in general not possible to define all concepts in a non-primitive manner. This is well described in [5]: "There are a large number of concepts that are unclassifiable by virtue of being natural kinds. The problem

is exacerbated by a large number of "fake" primitives, concepts which are primitive only because their definitions cannot be expressed in the restricted language. Since these reduce the utility of classification, using classification's efficiency as the design criterion misplaces emphasis."

Hence, there are various impediments to fully exploit the reasoning strengths that Description Logics offer. This means that in practice, classification may be overlooked that could have been inferred if concept definitions would have been non-primitive. This paper describes the possibilities of using non-primitive concept definitions in the process of knowledge modeling. We describe a method, developed to increase the utility of classification, especially during the knowledge modeling process. This method aims at utilizing DL inference services for the detection of: duplicate concept definitions, underspecification, and actual limits of a DL-based representation. Duplicate definitions should generally not occur in a knowledge base, and underspecification may point at tacit (i.e. not represented by a concept definition) knowledge. Minimizing tacit knowledge will increase the possibilities for distinguishing concepts based on their definitions, and may improve the model by reducing the number of primitive definitions.

We describe the knowledge modeling process in Section 2 and then explain our method in detail in Section 3. Results of the application of the methods in a case study are presented in Section 4 and discussed in section 5. Section 6 concludes this paper.

2 The Knowledge Modeling Process

Medical TKBs have grown in size and complexity. This growth has been stimulated by the availability of computers and the potential of using medical TKBs for a wide range of purposes. The complexity has increased due to the possibility of using representation formalisms that allow for more elaborate specification of concept definitions. Medical TKBs evolved from simple taxonomies to semantic networks with (informal and formal) concept definitions. An example of such a system is SNOMED CT[1], a terminological system consisting of approximately 350,000 concepts. Maintenance of systems this large needs to be supported as much as possible in order to reduce modeling errors. To this end we have started a project aiming at assessing and improving the quality of Medical TKBs. Previously, we have discussed the possibilities of detecting inconsistencies in concept definitions [3], and others have focused on this issue as well (e.g. [7]). However, (logical) inconsistencies are not the only modeling error that can occur. Generally, a good terminological system should fulfill a number of desiderata [1]. A formal, concept-oriented approach to modeling terminological knowledge can largely contribute to fulfill a number of these desiderata, for example providing formal definitions and multiple consistent views. However, representation alone is not sufficient. To fully exploit the advantages of formally represented knowledge, services, such as inference services, are indispensible. Standard Description Logic inference services (satisfiability and subsumption testing) provide a solid basis for supporting

[1] `http://www.snomed.org/`

1. InfectiousDisease $\equiv$ Disease $\sqcap \exists$ involves Infection
2. LiverDisease $\equiv$ Disease $\sqcap \exists$ location Liver
3. ViralHepatitis $\equiv$ InfectiousDisease $\sqcap \exists$ location Liver $\sqcap \exists$ cause Virus
4. DuplicateViralHepatitis $\equiv$ LiverDisease $\sqcap \exists$ involves Infection $\sqcap \exists$ cause Virus
5. PrimViralHepatitis $\sqsubseteq$ InfectiousDisease $\sqcap \exists$ location Liver
6. PrimDuplicateViralHepatitis $\sqsubseteq$ LiverDisease $\sqcap \exists$ involves Infection
7. ViralHepatitisTypeA $\equiv$ InfectiousDisease $\sqcap \exists$ cause HepatitisAVirus $\sqcap \exists$ location Liver

Figure 1: Examples of primitive and non-primitive concept definitions

knowledge modeling, but do not support all of the modeling process. For example, these services do not by default contribute to ensuring definition of non-redundant, unambiguous definitions that explicitly and maximally capture the semantics. However, although this is not supported directly, it is possible to utilize DL inference services for tasks such as ensuring definition of non-redundant, unambiguous definitions.

In the next section we focus on methods that utilize DL inference services for detection of concepts that are inadvertently defined more than once, and concepts that may not be exhaustively defined. With exhaustive definitions we mean definitions that maximally capture the semantics of the defined concept. Duplicate definitions should be prevented, as they introduce redundancy into the knowledge base. The motivation for focusing on detection (and reduction) of non-exhaustive definitions, is that non-exhaustively defined concepts may point at tacit (i.e. not represented by a concept definition) knowledge. Minimizing tacit knowledge will increase the possibilities for distinguishing concepts based on their definitions, and may improve the model by reducing the number of primitive definitions, supporting among others more elaborate classification.

3 Use of non-primitive Definitions

An essential feature of Description Logics is the distinction between definitions that state only necessary conditions (which we refer to as primitive definitions) and definitions that state necessary and sufficient conditions (non-primitive definitions). Non-primitive definitions facilitate the inference of classification (subsumption) of concepts. These definitions also make it possible to detect equivalent concepts (which actually boils down to mutual subsumption). For example, given the definitions 1 and 2 from Figure 1, equivalence of the concepts ViralHepatitis and DuplicateViralHepatitis, defined in 3 and 4, can be inferred. However, if these concept definitions would have been primitive, as in definitions 5 and 6, equivalence would not have been detected, and duplicate concepts would remain undetected. Furthermore, the concepts defined in 5 and 6 contain tacit knowledge, as these definitions do not state the viral cause of the disease.

As the domain of medicine consists of many natural kinds, for which no necessary and sufficient conditions exist, many disease concepts in Medical TKBs can only be

defined in a primitive manner. As a result of this, much of the inferential potential is lost, as the example above demonstrates. Another example would be missed classification. For example, given definition 7, ViralHepatitisTypeA would be rightly classified as a ViralHepatitis, but not as a PrimViralHepatitis, although it should be.

Although it is inevitable to have many primitive definitions in a Medical TKB, it makes sense to use the inferential powers of DL reasoners in the modeling process by stating, as an assumption, non-primitivity of all relevant concept definitions.

The first step is to determine which concepts in a knowledge base might have duplicate definitions or contain tacit knowledge that needs to be made explicit. Generally, medical TKBs consist of various "modules" that are used to define the concepts in the primary category of interest. Such TKBs can be regarded as a collection of subtrees, where the roots of these subtrees can be for example Diseases, Anatomical Components and Micro-Organisms. The Micro-Organisms subtree is used in the definitions of etiology of diseases, as is also shown in definition 7 in Figure 1. As a TKB is generally focused on one subtree, in this case diseases, it may be expected that the other modules are far from exhaustively defined. This will result in many equivalent concepts, not only in the respective subtrees, but also in the Diseases subtree, as concepts in this subtree are defined using concepts that were considered equivalent. There are two options in such a situation. The first option is to initially focus on the respective subtrees (such as Micro-Organisms), in order to find duplicates and make tacit knowledge explicit. The second option is to leave these subtrees out of account (i.e. treat the concepts as base symbols), and focus on the subtree of primary interest, which contains disease concepts in this example.

The next step is to find all concepts that have a definition of the form: $B \sqsubseteq A$, where A is a concept name. There is no use in changing these definitions to non-primitive definitions, as this will provide a trivial equivalence ($B \equiv A$). These concepts may be expected not to represent duplicate definitions (as they have been explicitly defined), but B either represents a natural kind or is underspecified. As concepts of this form are easily recognizable, they can be studied separately. The last step is to redefine all relevant concepts (e.g. disease concepts) that are not of the form $B \sqsubseteq A$ as non-primitive.

When the TKB has been altered according to the steps mentioned above, it can be classified with a DL reasoner. This classification will result in clusters of equivalent concepts. These clusters then have to be analyzed by hand. This analysis will provide two types of outcomes. First, a set of concepts that form duplicate definitions, which were previously undetected due to their primitive definitions. Second, a set of concepts for which the differences among them do exist, but are not represented. In the latter case, which we will refer to as underspecification, the knowledge base can potentially be improved by making explicit the implicit knowledge that distinguishes one concept from another. If this distinction can not be made explicit, this can either be caused by the lack of characteristic features of the concept (i.e. it is a natural kind), or by limitations of the used DL. We have previously performed a study on the use of terms indicating the need for certain constructors [2]. The above-mentioned method will contribute to gaining insight into the need for specific constructors by making the needs more precise and related to a specific TKB.

Cachexia ≡ Medical_Diagnosis ⊓ ∃ involved_system Metabolic_system
Starvation ≡ Medical_Diagnosis ⊓ ∃ involved_system Metabolic_system
Hypermagnesaemia ≡ Metabolic_Disorder ⊓ ∃ involved_component Body_fluids
Hypomagnesaemia ≡ Metabolic_Disorder ⊓ ∃ involved_component Body_fluids
Hypercalcaemia ≡ Metabolic_Disorder ⊓ ∃ involved_component Body_fluids
Hypocalcaemia ≡ Metabolic_Disorder ⊓ ∃ involved_component Body_fluids

Figure 2: Examples of equivalent concept definitions. Six concepts are defined, but classification results in two different concepts: Cachexia and Starvation as one concept, and Hypermagnesaemia, Hypomagnesaemia, Hypercalcaemia, and Hypocalcaemia as another concept

4 Results of two Case Studies

The approach is applied to a medical TKB on Reasons for Admission in Intensive care (DICE) [4] and on the Foundational Model of Anatomy (FMA)[2]. The DICE knowledge base, which is under development at the institution of the authors, contains about 2500 concepts, of which 1456 Reasons for Admission. Reasons for Admission comprise both diseases and procedures that require intensive care and monitoring of patients. We have applied the above-mentioned methods to detect duplicate definitions in the system, and to determine possibilities to improve modeling by reducing underspecifications. As we are aware of underspecification in domains other than Reasons for Admission (such as anatomy, and etiology), we have limited our evaluation to the Reasons for Admission taxonomy.

The case study on FMA has been performed mainly to determine to what extent the findings in DICE were knowledge-base-specific. FMA, developed by the University of Washington, provides about 69000 concept definitions, describing anatomical structures, shapes, and other entities, such as coordinates (left, right, etc.).

4.1 Results of the Case Study on DICE

The DICE knowledge base, which was implemented as a simple TBox with an empty ABox, was represented using Knowledge Representation System Specification (KRSS) syntax [6]. This made it straightforward to discern primitive definitions from non-primitive ones, simply by performing a text-based search in the KRSS file. Replacement of primitive definitions by non-primitive definitions for appropriate concepts out of 1456 Reasons for Admission, resulted in 108 (7%) concept definitions that were primitive, and 1348 (93%) concept definitions that were non-primitive. As was explained in Section 3, all primitive definitions were of the form B ⊑ A, for example: Eclampsia ⊑ Hypertension_induced_by_Pregnancy.

RACER[3] was used to classify the resulting TBox. This resulted in 24 unsatisfiable concepts, which we will discuss further in Section 5. Of the remaining 1432 satisfiable concept names (in the Reason for Admission module), 1160 (81%) had a unique defini-

[2] http://sig.biostr.washington.edu/projects/fm/

[3] http://www.sts.tu-harburg.de/~r.f.moeller/racer/

Table 1: Results of detection of equivalently defined concepts in the Reason of Admission module of DICE. The first column shows the size of clusters, the second column the number of clusters with the specified size, the last column shows the total number of concepts in the clusters of the specified size (i.e. the product of cluster size and number of clusters).

# equivalent definitions	**# clusters**	**# concepts in clusters**
2	60	120
3	23	69
4	8	32
6	2	12
7	4	28
11	1	11
	98	272

tion, and 272 concepts (19%) were equivalent to one or more other concepts. As shown in Figure 2, classification will render multiple concepts with equivalent definitions as one concept. The 272 concepts could be traced back to 98 definitions that were used twice or more, as is shown in Table 1. There were 60 tuples of equivalent definitions (such as Cachexia and Starvation), and 1 cluster with 11 concepts. This last cluster contained concepts as diverse as Water_Depletion and Familial_Periodic_Paralysis.

4.2 Result of the Case Study on FMA

The FMA knowledge base, which is implemented as a frame-based model in Protege[4], has been migrated to DL, where specified slot-fillers in the frame-based representation were interpreted as existentially quantified roles. This simple TBox, with an empty ABox, was represented using KRSS syntax. Replacement of primitive definitions by non-primitive definitions for appropriate concepts, resulted in a DL-based representation of all of FMA that contained about 50% primitive and 50% non-primitive concept definitions. We were not able to classify the full TBox with RACER, probably because of the use of roles and their inverses (e.g. part_of and part), leading to cyclic definitions. Because of this, we limited the case study to "Organs". Of the 3826 concept definitions, 2659 (69%) were non-primitive, and 1167 (31%) were primitive. Classification with RACER resulted in 3323 concepts (87%) that had a unique definition, and 503 concepts (13%) that were equivalent to one or more other concepts. These 503 concepts could be traced back to 160 definitions that were used twice or more, as is shown in Table 2. There were 106 tuples of equivalent definitions, and 1 cluster with 54 equivalent concepts. This cluster contained a variety of ligaments of joints, such as Interosseous_ligament_of_carpometacarpal_joint and Palmar_ligament_of_left_fifth_carpometacarpal_joint.

[4] `http://protege.stanford.edu/`

Table 2: Results of detection of equivalently defined concepts in the Organ module of FMA. The first column shows the size of clusters, the second column the number of clusters with the specified size, the last column shows the total number of concepts in the clusters of the specified size (i.e. the product of cluster size and number of clusters).

# equivalent definitions	**# clusters**	**# concepts in clusters**
2	106	212
3	33	99
4	10	40
5	4	20
6	3	18
7	1	7
$\geq$8	3	107
	160	503

4.3 Explanation of equivalence

As described in Section 3, there can be various explanations for concept equivalence. Concepts can be actually duplicately defined, but can also be underspecified. This underspecification is inevitable when concepts are natural kinds, or when concept properties can not be expressed due to limits of the used DL. Avoidable underspecification indicates tacit knowledge, that could be made explicit by making definitions more exhaustive. Although a full evaluation of equivalent definitions has to be performed, a first study provides markable results.

In DICE, only 4 tuples of concept definitions were found that are potential duplicates, but this needs to be discussed with domain experts. Examples of such concepts are "reconstruction of artery" versus "arterial angioplasty" and "biliary drainage" versus "drainage of biliary duct".

Apart from these potentially true duplicates, all equivalent concepts differ in meaning in a way that is not represented in the knowledge base. In DICE, a small number of natural kinds was found, which were to a large extent syndromes and/or eponyms. Examples of these are "Adult Respiratory Distress Syndrome", "Wilms' tumour", and "Wolff-Parkinson-White syndrome".

Both DICE and FMA originally have a frame-based representation, and both have been migrated to DL in order to be able to perform the experiments described. In DICE, a small number of concepts was found that explicitly mentioned negation, which can not be represented using frames. Examples of these are "bleeding" versus "non-bleeding" and "obstructive" versus "non-obstructive". This difference could be explicitly represented using a DL that allows for negation.

The vast majority of concepts that were defined as primitive or that were non-uniquely defined, demonstrated underspecification that seemed to be relatively easy to avoid. This means that it is possible and appropriate to make definitions more

exhaustive by adding conditions.

Possible improvements to DICE can be determined by studying equivalent concepts. For example, equivalence of hypocalcaemia and hypercalcaemia can be resolved by making explicit the level involved: "below normal" resp "above normal". Equivalence of hypercalcaemia and hypermagnesaemia is explained by the lack of specification of the involved chemical elements (calcium resp magnesium). These examples demonstrate required extensions to the knowledge base, as chemical elements and levels are currently not defined in the knowledge base. There were however also many concepts that can be refined using concept and roles that are already available in the knowledge base. For example "Meningococcal meningitis" was defined as a "Bacterial meningitis", without mention of a relation with a concept "Meningococcus" through an "etiology" role. Hence, making such a definition more exhaustive is not only straightforward, it is even required, if one wants to classify meningococcal meningitis as a disease that is caused by meningococcus. Making concept definitions more exhaustive using readily available concepts and roles, also seemed possible in FMA. For example, "Synovial tendon sheath of flexor hallucis longus" and "Synovial tendon sheath of tibialis anterior", can be distinguished by explicitly relating them to "flexor hallucis longus", and "tibialis anterior", respectively.

5 Discussions

The two case studies demonstrate the feasibility and usability of our approach. In order to assess the overall applicability of the approach, it is useful to look further into the peculiarities of the knowledge bases used in the case studies. We will discuss these below. Thereafter, we will shortly discuss an alternative approach that could be used instead of our method, namely structural subsumption.

5.1 Modeling Issues

The knowledge bases that have been studied exhibit a number of properties that render them suitable for the method described in this paper. Both DICE and FMA are represented as simple TBoxes. This implies that no atomic concept occurs more than once as left-hand side, and the left-hand side of all axioms are atomic concepts (so no arbitrary concept expressions are allowed on the left-hand side). Moreover, the DICE TBox is acyclic, meaning that no concept name is defined with reference to itself (such as for example: Human $\equiv$ Animal $\sqcap \forall$ hasParent Human). The FMA TBox contains cycles, caused by the use of roles and their inverses (e.g. part_of and part). The similarity and simplicity of both knowledge bases can be explained by the fact that the DL-based representations are the result of a migration process from a frame-based representation, as described in [3].

The FMA TBox was coherent, and the DICE TBox contained only a small number of unsatisfiable concepts (due to the migration process). Having a minimum of unsatisfiable concepts is important as unsatifiability "propagates" over existentially quantified roles. Assume that B is an unsatisfiable concept, then C, defined as C $\sqsubseteq$ A $\sqcap \exists$ R B, will also be unsatisfiable.

In order to minimize the number of unsatisfiable concepts, no disjointness between concepts was explicitly stated. This can be explained by the following example. Suppose the original knowledge base contains two primitive definitions: $C_1 \sqsubseteq A \sqcap \exists R\ B$, and $C_2 \sqsubseteq A \sqcap \exists R\ B$, and C_1 and C_2 are stated to be disjoint. Applying our method would change these definitions to non-primitive definitions, which would render C_1 and C_2 equivalent. But as they are also disjoint, each would be inferred to be unsatisfiable.

The DICE TBox is defined using the language $\mathcal{ALCQ}$, the FMA TBox can be expressed with $\mathcal{ALCI}$. This means that for example role hierarchies and transitive roles are not used in these TKBs. Actually, DICE was modeled using Structure-Entity-Part (SEP) triplets, described in [8], in order to prevent the use of transitive roles and role hierarchies.

It needs to be determined whether the method is also useful for more complex knowledge bases. Issues that increase the complexity of knowledge bases are the use of a more expressive language, cyclic definitions, use of concept inclusion axioms with concept expressions on the left-hand side (instead of only atomic concepts), and allowing multiple definitions of a concept.

5.2 Alternative Approach: Structural Subsumption

There are two reasons for discussing structural subsumption as an alternative approach. The first reason is the fact that the case studies involved relatively simple knowledge bases. The second reason was that a superficial inspection of equivalent definitions of concepts indicated that most of them were not only logically equivalent (as definitions 3 and 4 from Figure 1), but even structurally equivalent, as shown in Figure 2. Structural subsumption could prove useful for knowledge bases for which the computational cost of classification is too high, such as for example the complete FMA. The advantage of using a structural subsumption algorithm is that it is generally cheaper in terms of computational costs, but it has the drawback that it is not complete.

6 Conclusions

We have applied the inferential powers of DL reasoners to detect concepts that are equivalently defined within a knowledge base. In order to be able to find such concepts, we have considered relevant concept definitions as non-primitive. This results in clusters of concepts which have equivalent definitions.

Two case studies show that the size of such clusters varied mainly from 2 to 7. For the vast majority of concept definitions that turned out to be equivalent it is possible to make them more exhaustive by adding conditions that distinguish between them. For a minority of the equivalent concepts there seemed to be no possibilities of making the definition more exhaustive, as these concepts represented natural kinds, or could not be defined due to limits of the underlying representation. The case study on DICE revealed only a few duplicate definitions in the knowledge base.

Overall, it can be concluded that applying the methods described in this paper, contributes to gaining insight in tacit knowledge, which is unrepresented in a knowledge base. Making this knowledge explicit by means of refining concept definitions improves the knowledge base, and results in more exhaustive concept definitions. However, it can not be guaranteed that these more exhaustive definitions will now provide both necessary and sufficient conditions. Therefore, it needs to be determined whether this method can result in an actual decrease in the number of primitive concept definitions in knowledge bases, which would increase the powers of inference based on the knowledge base. The successful application of our method to two knowledge bases in the field of medicine, makes it likely to be applicable to other domains as well.

References

[1] J. J. Cimino. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine*, 37(4-5):394–403, 1998.

[2] R. Cornet and A. Abu-Hanna. Usability of expressive description logics – a case study in UMLS. *Proc AMIA Symp*, pages 180–4, 2002.

[3] R. Cornet and A Abu-Hanna. Using description logics for managing medical terminologies. In M. Dojat, E. Keravnou, and P. Barahona, editors, *9th Conference on Artificial Intelligence in Medicine in Europe, AIME*, pages 61–70, Protaras, Cyprus, 2003. Springer.

[4] N. F. de Keizer, A. Abu-Hanna, R. Cornet, J. H. Zwetsloot-Schonk, and C. P. Stoutenbeek. Analysis and design of an ontology for intensive care diagnoses. *Methods of Information in Medicine*, 38(2):102–12, 1999.

[5] Jon Doyle and Ramesh Patil. Two theses of knowledge representation: Language restrictions, taxonomic classifications, and the utility of representation services. *Artificial Intelligence*, 48(3):261–298, 1991.

[6] PF Patel-Schneider and B Swartout. Description-logic knowledge representation system specification from the krss group of the arpa knowledge sharing effort. Technical report, KRSS Group of the ARPA Knowledge Sharing Effort, 1 november 1993 1993.

[7] E. B. Schulz, J. W. Barrett, and C. Price. Semantic quality through semantic definition: refining the read codes through internal consistency. *Proc AMIA Annu Fall Symp*, pages 615–9, 1997.

[8] S. Schulz, M. Romacker, and U. Hahn. Part-whole reasoning in medical ontologies revisited–introducing sep triplets into classification-based description logics. *Proc AMIA Symp*, pages 830–4, 1998.