With Semantic Analysis from Noun Phrases to SNOMED CT and Classification Codes

H. R. Straub, M. Duelli, Semfinder AG, Kreuzlingen, Switzerland

Abstract

For the classification of freely formulated diagnostic noun phrases in a fully automated way, the entire semantics of the diagnoses must be represented in a machine-readable form. Any problems resolved in this process re-emerge when assigning SNOMED terms to texts from patients' records or when assigning classification codes to SNOMED terms. We discuss the challenges and conditions associated with resolving these issues and the possibility of allocating SNOMED terms and classification codes simultaneously.

1. Introduction

In order to exchange and compare data in the medical field, standardization is essential. While it is not easy to agree on an external data standard, the problems become even more complex when we try to find a standard for meaning (semantic interoperability). Semantics, i.e. the meaning of words, can't be measured and formalised quite so simply as when dealing with technical constructs. The meaning of words doesn't seem to be merely dependent on the words themselves, but also on the context and may change from speaker to speaker and from time to time. Nevertheless we would like to find a standard for semantic interoperability and believe that this should be possible, at least in a defined scientific specialist area such as medicine.

As an extensively developed standardized terminology, SNOMED CT would seem to be the most promising option in terms of achieving an interoperability standard of this kind. There are 2 practical questions if SNOMED CT is to be used in hospitals on a day-to-day basis:

A: How can we assign the correct SNOMED terms to the medical data, and in particular to diagnoses texts in patient records?

B: How can we obtain the accurate classification codes (ICD-10 and procedure codes) which we will still require in the future[4] from the SNOMED terms?

Problem A would seem to be resolved with the precoordinated terms available in SNOMED CT and Problem B by the use of mapping tables, but we still need to demonstrate that both problems can be resolved with a degree of certainty in a practicable way on a day-to-day basis.

Our team has created a system which is used in daily routine in hospitals in Germany to automatically assign ICD-10 codes to free diagnostic terms. To this end we had to work out how meaning (semantics) in free text can be recognised and formalised (Problem A) and what information we require to allocate classification codes with sufficient certainty (Problem B). The paragraphs below will discuss some of the conditions which are required in order to resolve both of these problems.

2. From freely formulated diagnoses to ICD-10 codes

In the automated coding system, the content, i.e. the semantics of the freely formulated diagnosis, must be presented in a form which can be read by both humans and computers (Fig. 1). This must be in a highly structured form to be read by machine. In the same way, the classification system (ICD-10 in this case) must be analysed semantically. Each individual class (ICD-10 code) with all inclusiva, exclusiva and comments, along with the additional explicit German coding rules[1] in Germany, must be understood and be able to be represented in the specified computer-legible form. If the specified formal representation is genuinely capable of representing both the semantics of the free text and of the ICD-10 code in a clear and computer-legible way[8], the coding process from Fig. 1 can be performed fully by machine. Unlike traditional coding tools, which offer the user a selection of possible codes for his "search text", this system is able to make a definitive decision on coding (One Step Coding).

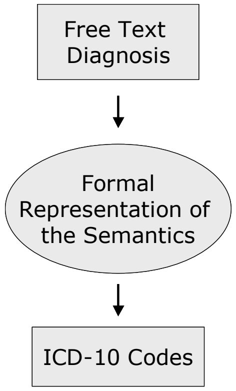

Fig. 1: Fully automated classification coding

Of course, the condition for such One Step Coding is that the text contains all the information needed for the selection of the specific ICD-10 code. If the text is too vague, additional information must be given. The outline of this interactive process is shown in section 3.3.

A classification as ICD-10 serves a particular purpose and cannot claim to be a general representation of the world. The noun phrases leading to the code are much closer to the information of the real world, but still very far away from its full information content. How perfect they are built, neither a classification, a terminology, a natural language nor any form of formal ontology can reach the full information content of reality. The chance that any given representation misses just that bit of information that another representation needs, remains always imminent. In our task of automated classification coding we therefore did not primarily search for the gold standard in the choice of the represented information items but for the most suitable method for presenting, comparing and extending the information of a given "language". Completeness was never a goal, but ease in dynamic modelling was.

The ICD-10 classification shows expressions like "NOS" (not otherwise specified), dagger/asterisk combinations (Example 2 below) and complex inclusion/exclusion constructs (Example 1). All these challenges point to general problems of semantic representation and are found in the representation of the input noun phrases as well. In section 3 I give a summary of the problems: While the dagger/asterisk combination can be seen as a combination of two axes (sections 3.1 and 3.2) and is easy to represent, the "not otherwise specified" corresponds to the absence of values of one (or several) specific axes (section 3.3) and can – with a clear identification of the qualities (axes) – also easily be represented. More difficult, however, are the complex inclusion-exclusion constructs (Example 1), especially when the sum/summands problem (section 3.4) must be solved.

Example 1 - Excludes, includes and combinations of several diseases resulting in one code:

I12	**Hypertensive renal disease** *Includes:* any condition in N18.- (renal failure) … … with any condition in I10 arteriosclerosis of kidney, … *Excludes:* secondary hypertension (I15.-)
I12.0	**Hypertensive renal disease with renal failure** hypertensive renal failure
I12.1	**Hypertens. renal disease without renal failure** hypertensive renal disease NOS

Example 2 - Asterisk/dagger combination:

"Candida pneumonia" is coded as:

B37.1(†)	Pulmonary candidiasis
J17.2*	Pneumonia in mycosis

3. Challenges in the automated coding process

3.1. Combinatorial explosion and use of thesauri

Medical diagnoses are known to be extremely varied[5] and contain information on different features which may be freely combined with each other. Each feature can have many synonyms and quasi-synonyms and be specified with varying degrees of precision (granularity). The expressions which can be used in this way combine explosively and as a result it is not possible to put up a list of all combinations in a thesaurus as a platform for code classification. It is more appropriate to select a formal representation of the semantics which has itself the ability to combine. The question is how many axes (dimensions, degrees of freedom) we need to offer for combination purposes. Our experience indicates that combinatorial explosion can only be resolved by using a system which a) offers an unlimited number of axes and b) is able to structure and encapsulate[8] the axes with reference to each other. This last requirement is essential in order to keep the system transparent and maintainable.

The SNOMED CT System is multi-dimensional (in postcoordinated form), but only has a limited number of axes. The additional precoordinated terms correspond to a thesaurus and, like any thesaurus, offer no guarantees against combinatorial explosion.

3.2. Granularity and number of axes

We also need to know how fine the granularity of the semantic system needs to be. Obviously, no system which can be created in practice can have unlimited granularity. Although it is very important that granularity should be as fine as possible to ensure the semantic power of a system (and thus its suitability as a platform for semantic interoperability), in practice no system can operate with an infinite number of values. It is more important to maximise the ratio of expressive power and expressive effort by careful structuring of information.

This subject is closely linked to the previous subject. The more axes are offered and the more logical the links between them, the less information needs to be packed on to one individual axis, reducing the need to differentiate the individual axis. The axes are then combined, which multiplies their expressive power and thus deals with the combinatorial explosion of real world ontology. A multi-axis system (multi-dimensional) is superior to a single-axis (one-dimensional, one-hierarchical) system, as well as a system with many axes is superior to one with only a few. Our system contains over 1000 axes (degrees of freedom) and can thus achieve a very fine representation which corresponds to a thesaurus with several billion entries. When dealing with so many axes, their relative positions are important. The way the axes are structured with respect to each other becomes crucial. It provides logical paths for the interpretation algorithms which is useful not only when constructing and maintaining the system, but also to ensure fast performance of the system in use.

3.3. Incomplete information in texts

What can be done when the free text diagnosis is incomplete for classification purposes? This happens frequently. For example, let us consider a case of meniscus damage, which may be an acute trauma or a chronic degenerative lesion and will be coded accordingly using an ICD-10 code from a completely different section. The doctor may, however, merely write "partial lesion of posterior horn of left medial meniscus" on his notes. Despite the precise reference to location, the diagnosis cannot be assigned with sufficient certainty to a particular classification section. Practice shows that such cases occur frequently as the doctor making the notes does not know what criteria are most important for classification purposes or supposes that the corresponding information is self-evident in his context.

There are two possible options for coding incomplete texts: either just the code provided by the coding system as a default value is selected – thus most certainly leading to incorrect coding in many cases - or the coding program offers an interactive solution (figure 2).

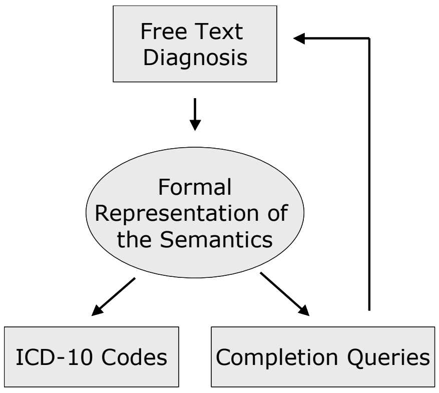

Fig.2: An interactive coding system

In the interactive solution, it must be possible to clearly identify in the formal representation not the missing words or concepts, but the missing feature (quality) which will be asked for by the completion question. Such identification is only possible if the feature itself is clearly defined as en entity, best in the form of a degree of freedom in its own right in the semantic representation, i.e. as an own axis or dimension in the concept architecture[8]. This provides further proof of the need to supply as many axes/dimensions/degrees of freedom as possible. If this were not the case, the absence of a quality would not be readily and directly identifiable, especially if its values could each be represented by a bunch of possible synonyms, and values as well as synonyms could change its meaning depending on the context. The algorithms which would be needed to work with such value lists instead of qualities (axes) would not only be slower, but would also be much more difficult to maintain.

However, by having a semantic representation model in which axes can be identified in an easy way, it is a simple matter to automatically identify missing qualities in the formulation of diagnoses and offer the user an interactive list from which to select possible values (Fig. 2).

3.4. Sum/summands problem (optional part-of)

A diagnosis does not always lead to one code and classification entry. It is entirely possible that a diagnosis may need to be assigned to two codes at the same time. Dagger and asterisk coding in the ICD-10 is typical of such cases – but is by no means the only type of multiple coding method.

The reverse is also possible, i.e. two ailments afflicting one patient will be assigned to just one ICD-10 code. This is because illnesses are not indivisible entities which can be isolated; instead, they interact in a complex system of causality and coincidence. It is often possible for there to be codes for the individual illnesses (summands) and for the overall picture (sum). It is usually possible to express more details with the summand codes than with the sum code, but the sum code also shows the coincidence of the individual illnesses and thus complies much better with the notion of a classification which assigns a case precisely to one place and class.

The sum/summand problem has been explained in greater detail using the example of arterial hypertension and its secondary diseases[9]. I should like to reiterate at this juncture that the sum/summand problem in diagnoses has little to do with the well-known "part-of" relationship, as is customary when describing anatomy. It is absolutely necessary to distinguish between an *obligatory* and an *optional* (contingent) “part-of”. In the obligatory “part-of" relation as seen in anatomy, it is always possible to draw a conclusion from the finger to the hand, and from the hand to the fingers. Such deductions are by no means inevitable when considering diagnoses from the sum/summands viewpoint: the sum of "coronary heart disease" cannot be used to deduce the optional summand "myocardial infarction", and the summand "osteoarthritis of the knee" cannot be used to deduce the possible sum "generalised osteoarthritis". The relationship between summand and sum is much looser (= optional "part-of") and needs to be specified individually in every single case. It should also be borne in mind that a summand diagnosis can be part of a sum diagnosis as well as a complete diagnosis in its own right, quite unlike in anatomy, where the finger does not exist in isolation from the hand.

Schulz et al[6] discuss the problems arising from the sum/summand situation with an examplary procedure. The same challenges for the concept representation are found in his example of procedures as in ours of diagnoses.

The sum/summand problem of diagnoses and procedures needs to be approached in logical computer terms quite differently than the “part of” relationship in anatomy. While the obligatory “part of” in anatomy barely poses any problems for automated diagnostic and procedure coding, the sum/summand problem is the real challenge.

4. Can terminologies replace classifications?

As terminologies have finer granularity than classifications, the assumption is that they already contain all the information which is expressed by the classification and can therefore replace classifications. However, this is only true to a certain extent. This is because the field of diagnoses is extremely complex and its system forms a hierarchy only in a very imperfect fashion.

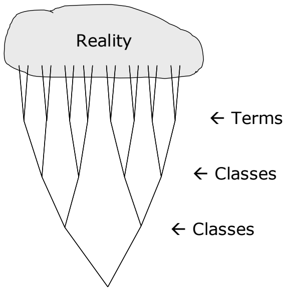

Fig. 3: Hierarchical model of terms and classes (idealisation)

The model in Fig. 3 corresponds to a very simplified view of things. The respective hierarchical system classifying objects of reality in terms and classes is not compulsory and can come about in different ways. This means that terms can be assigned to classes in more than one way. Fig. 4 shows this situation in schematic way. In reality, the situation is even more complex: Further problems are that a) terms not only need to be classified in one hierarchy as shown, but in several simultaneously (Sections 3.1. and 3.2.), and that these hierarchies interact in a complex way, b) a number of terms combine to form a diagnosis, i.e. several points need to be considered simultaneously when representing as a class and c) unlike in Fig. 4, the terms are not fixed points but are already classifying simplifications which combine a number of different instances of reality, a combination which could always be made in a different way, as practice shows.

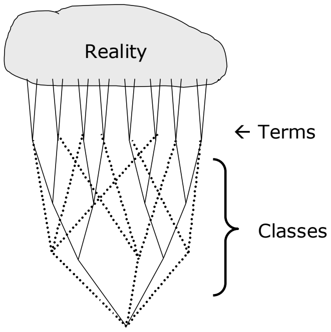

Fig. 4: Alternative class hierarchies

The information loss when we make the transition from terms to classes is both inevitable and desirable. When we classify, we do *not* actually *want* to know every detail of the instances so that we can have a clear overview of the cases in hand. While information loss is natural, selecting the information which will be lost is not a natural process, but depends on the intended purpose of the classification system. For all of these reasons information on classification cannot automatically be derived from terminology.

Classifications have a value all of their own for the reasons described above[4]. It should also be quite clear that, unless we make specific interactive completions, it is not possible to ensure mapping of SNOMED to ICD-10 without losing precision and this is not tolerable when applying ICD-10 in the DRG[1,3] field for example.

5. Semantic free text analysis to allocate SNOMED CT and ICD-10 Codes simultaneously

Automated semantic noun phrase analysis (Fig. 2) for the allocation of ICD-10 codes is successfully used in daily hospital routine. When analysing the noun phrases, their semantics need to be fully understood, which means that the semantic representation of the diagnoses not only contains the concepts relevant for the ICD-10, but all semantic dimensions (qualities) identifiable in the diagnostic texts as well as the logic structure of the resulting constructs (section 3.2). Once the full semantics of the texts are represented, they can be translated to other forms of information, be it to a coarse granular classification (ICD-10) or to a more precise terminology (SNOMED CT). The problems of these translations are outlined in section 3 and are not greater when translating to SNOMED CT than when translating to ICD-10. There is no reason not to derive SNOMED CTs from the internal representation arising during ICD coding and not to do both at the same time.

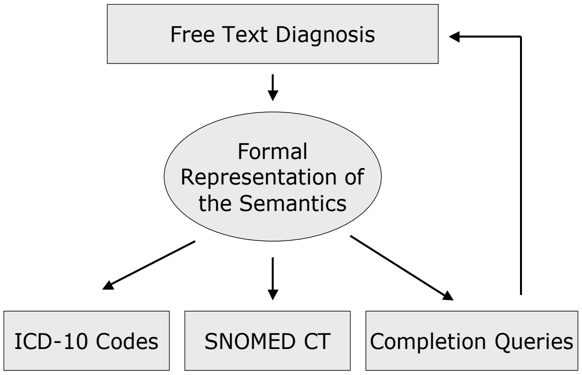

Fig. 5: Simultaneous assignment of SNOMED and ICD codes

Users in hospitals who will also have to allocate ICD-10 codes in future will be pleased that they do not have to do the same job twice. Combining the two coding systems offers the advantage that SNOMED CT coding could benefit considerably from ICD-10 coding. The challenges described in Section 3 are known and will be overcome by technology which is already available. Incomplete expressions are completed by the interactive query routine and for the classification a comprehensive finalisation process takes place, which will improve the the selection of SNOMED terms describing the case, too.

The internal representation which we use is not a standard, but a method with a net of concepts and qualities which can be extended dynamically at any time. In other words, the representation is open to many points of view and different takes on reality. This includes SNOMED CT like any other standard terminology. SNOMED CT, however, has a quite different function. It defines a fixed standard which can be used to communicate from clinic to clinic and from country to country. Both functions, the flexible method and the fixed communication standard, complement one another.

6. Conclusions

Practical aspects of implementing and using SNOMED CT in hospitals must not be underestimated.

SNOMED CT as a standard must be fixed, complete and unchangeable. This restricts its use as an open and dynamic semantic platform.

The solutions to problems discovered with the automated ICD-10 coding system may prove useful.

It is possible to simultaneously assign SNOMED terms and ICD-10 codes.

7. References

[1] Deutsche Krankenhausgesellschaft DKG, InEK gGmbH. *Deutsche Kodierrichtlinien, Version 2006*.

[2] Deutsches Institut für Medizinische Klassifikation und Information DIMDI. *ICD-10-GM Version 2006*.

[3] Fetter RB, Brand A, Dianne G (eds.). *DRGs, Their Design and Development*. Health Administration Press, Ann Arbor, 1991.

[4] Projektgruppe "Standardisierte Terminologien in der Medizin" (STM) der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS). *Positionspapier zur "Systematized Nomenclature of Medicine – Clinical Terms" (SNOMED CT) in Deutschland*. http://www.gmds.de/pdf/publikationen/stellungnahmen/Positionspapier.pdf (last accessed July 13, 2006).

[5] Rector AL. Clinical Terminology: Why is it so hard? *Methods of Information in Medicine.* 1999. 38(4-5): pp. 239-52.

[6] Schulz S, Hahn U, Rogers J. Semantic Clarification of the Representation of Procedures and Diseases in SNOMED CT. *Proceedings of MIE 2005*: pp .773-778.

[7] Spackman KA, Dionne R, Mays E, Weis J. Role Grouping as an extension to the description logic of ONTYLOG, motivated by entity modelling in SNOMED. *Proceedings of AMIA 2002:* pp. 712-716.

[8] Straub HR, Frei F, Mosimann H, et al. Simplified Representation of Concepts and Relations on Screen, *Proceedings of MIE 2005:* pp. 799- 804.

[9] Straub HR, Duelli M, Mosimann H, et al. From Terminologies to Classifications – the Challenge of Information Reduction. *Proceedings of the European Federation for Medical Informatics Special Topic Conference, Timisoara, Romania, 2006*: pp 341-352. See also: http://www.semfinder.ch/media/information_reduction.pdf (last accessed July 14, 2006)

Address for Correspondence

Hans Rudolf Straub, Semfinder AG, Hauptstrasse 23, 8280 Kreuzlingen, Switzerland. straub@semfinder.com