The New Navigators: from Professionals to Patients
R. Baud et al. (Eds.)
IOS Press, 2003

A Frame-based Representation of ICD-10

Paul Fabry[a], Robert Baud[a], Patrick Ruch[c], Pierre Le Beux[b], Christian Lovis[a]

[a]*Division d'Informatique Médicale, Hôpital Universitaire de Genève, Suisse*
[b]*Laboratoire d'Informatique Médicale, Faculté de Médecine, Rennes, France*
[c]*Laboratoire d'Informatique Théorique, EPFL, Lausanne, Suisse*

Abstract :

Physicians are required to code information concerning a patient's stay in order to measure the medical activity in hospitals. They use the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10). Coding is usually performed manually and computerized tools may be useful in speeding up and facilitating the tedious task of coding patient information. The aim of this work is to build a surface semantic model of ICD-10 in order to ameliorate a coding help system.
Methods: this work was focused on chapter XI of the ICD-10, Diseases of the Digestive System. Each term from both analytical and alphabetical indexes about this chapter were submitted to a morphological analysis in order to extract the medical concepts within. After a statistical analysis of these concepts and the way they connect themselves, a semantic model based on a "semantic frame" approach was built.
Results: although this model could represent a reasonable amount of medical knowledge within chapter XI of the ICD-10 in a quite satisfactory way, it shows lack of efficiency for some other chapters.
Conclusion: difficulties have to be overcome when modelling a classification meant for manual utilisation, and a lot of work still has to be done to obtain an effective coding help system using the ICD-10.

Keywords:

Coding; ICD-10; Knowledge Representation; Semantic Model; Frames

Introduction

The International Statistical Classification of Diseases and Health Related Problem published by the World Health Organisation was primarily designed to index causes of death and morbidity for statistical and epidemiological analysis. Its Tenth revision (ICD-10) has been released in 1993 [1].
This classification is an important source of medical knowledge, furthermore many countries including Switzerland now use it as the basic information source for tools measuring medical activities and planning health costs. Diagnoses are now systematically encoded. This tedious task, coping with ICD-10 two large volumes, is either handled by physicians, or more often by coding clerks, on the basis of patients' discharge summaries. Coding is not the main concern of physicians; therefore the data provided may be short of accuracy [2]. The weakness of manual coding added to an increasing demand for encoded data call for computer assisted coding tools which would be useful in speeding up this process and providing more reliability. Some already exist and one can roughly tell apart ICD-10 "browsers" allowing users to search a code with different criteria [3], from tools based on natural language query analysis [4].
The aim of this work is to suggest a way to represent knowledge within ICD-10 terms through a Frame-based semantic model and to evaluate its coverage and consistency.

Material and methods

The ICD-10 classification

A classification consists of a set of terms or elements ordered according to some specific criteria [5]. In ICD-10, these elements are health related problems. This classification is represented by a set of alphanumerical codes, each related to a main term and including several "secondary" include terms.
In addition, ICD-10 has an alphabetical index of the definitions (Volume III) that complements the analytical index (Volume I) where codes are ordered by chapters, themselves subdivided in blocks and sub-blocks. This hierarchy is continued by three-character codes (or category), subsuming four-character codes and sometimes five-character codes. In order to represent the content of the ICD-10 terms, relationships from a hierarchy position represented by a category to the main concepts present in the subjacent terms have been defined. They are:

1. A "Pathology" relationship where codes express different pathologies subsumed by the category.
2. A "Location" relationship where codes express different location of the category.
3. An "Association" relationship where codes express different complications of the category.
4. An "Aetiology" relationship where codes express different aetiologies of the category.

ICD-10 was conceived for manual exploitation. For numerous codes, terms include negations and relative complements e.g. expressions like "not elsewhere classified", relative to the practical use of codes. These expressions facilitate the manual course of ICD-10 but add imprecision in terms, hindering their automatic treatment.
In a practical way, ICD-10 includes 21 chapters with more than 18,000 codes and nearly 50,000 terms in both analytical and alphabetical indexes. This work is based on a Relational Database version of ICD-10 [6] and is voluntary restricted to one specific chapter of the French version.
Chapter XI: "Diseases of the Digestive System" has been chosen because it is frequently used for coding and it share a common structure with several others chapters. This chapter include 476 elements: 71 categories subsuming 405 codes. 2958 terms, from both analytical and alphabetical indexes have been kept after the elimination of duplets.

The UMLS Semantic Network [7]

Initiated in 1986 by the National Library of Medicine, the Unified Medical Language System has for its main objective to collect, in a rational way, the medical knowledge contained in numerous classifications throughout the world. This knowledge is distributed in three "Knowledge Sources":

1. The Metathesaurus is a set of concepts (776 940 in the eleventh edition) used for indexing biomedical concepts and terms from many classifications and controlled vocabularies.
2. The SPECIALIST Lexicon offers morpho-syntactic information and sub categorisation frames on more than 130,000 terms used in medical language.
3. The Semantic Network provides a consistent categorization of all concepts represented in the UMLS Metathesaurus through its 134 semantic types.

Knowledge extraction and analysis

First, each term undergoes an automatic morphological analysis, providing a sequence of morphemes. Some of these morphemes are manually brought together in order to keep a medical meaning. For example, "*gastroeosophageal*" will be split up into two distinct morphemes: "*gastro*" and "*oesophageal*", in return "*small intestine*" although this term includes two morphemes, will be kept together.
Each of these "medical morphemes" is either automatically coupled with the UMLS Semantic Type (ST) having the closest meaning via a lexicon [8], or manually defined for missing links in the lexicon. Three additional types are created in order to represent otherwise unclassifiable entities: negations, relative complements and proper nouns. Whenever possible, each morpheme is assigned to one of four main axes : *Pathology*, *Location*, *Association* and *Aetiology*, depending of the function of the morpheme within the definition.
A statistical analysis is performed on this data for determining axes' frequency, STs frequencies and distribution.

Model definition

Unlike former studies [9], [10], the intent of the authors is to represent the knowledge inside each distinct term rather than the whole ICD-10 hierarchy. With this objective in mind, a "semantic frame" approach is chosen. This kind of framework offer to decompose a knowledge domain to a set of frames, themselves defined by attributes or "slots" [11].
The working hypothesis is that a definition could be characterized by four axes: *Pathology*, *Location*, *Association* and *Aetiology*. Such a multi-axial representation already exists in SNOMED [12].
Each axe is represented by a frame in the model. Afterwards, each frame is attributed a set of slots with the help of the statistical analysis of STs distribution. Furthermore, the presence of negations, relative complements and proper nouns in some definitions requires to define others slots apart from the four main frames (Figure 1).

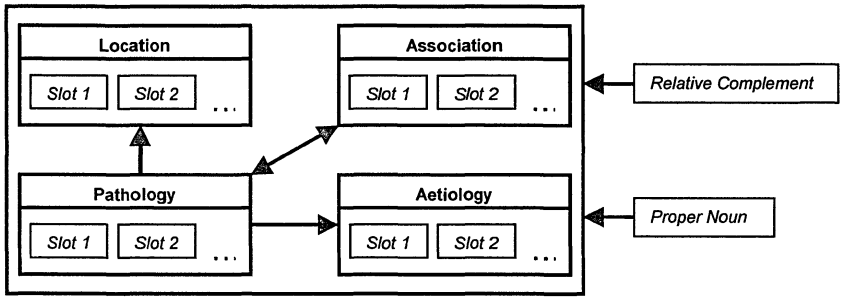

Figure 1: Semantic Model basic structure

Each definition of chapter XI is processed this way in order to evaluate coverage of the model.

Results

Data extraction and statistical analysis

The morphological analysis of the 2958 definitions has produced 9237 morphemes. 46 STs have been used for categorizing all the morphemes, apart from the three especially created. STs distribution is heterogeneous. 7 STs are sufficient to represent more than 75 % of the morphemes (Table 1), in return 22 STs are needed for categorizing 1 % of the morphemes. This ST's distribution seems to conform to a traditional Zipf's distribution [13]. ST T080: Qualitative concept is quite an imprecise concept; its importance outlines the difficulty of accurate morpheme categorization using the Semantic Network. Furthermore, the recurrence of relative complements within the definitions (ST: U001) is not negligible. Axes Pathology and Location are preponderant in categorizing the definitions (Table 2). A study of co-occurrences shows that the four axes are sufficient to entirely categorize about 75 % of the given terms.

Table 1: Main STs used for morphemes categorization.

ID	Definition	Number	%
T023	Body Part, Organ, or Organ Component	2744	29,7
T046	Pathologic Function	2363	25,6
T080	Qualitative Concept	712	7,7
T184	Sign or Symptom	354	3,8
U001	Relative Complement	314	3,4
T020	Acquired Abnormality	269	2,9
T047	Disease or Syndrome	255	2,8
	TOTAL	**7011**	**75,9**

In return, numerous terms include morphemes apart from the four axes. These morphemes are mainly relatives to proper nouns, negations and relative complements.

Table 2: Axes frequencies within the definitions (N = 2958)

Axes	Number	%
Pathology	2931	99,09
Location	2664	90,06
Unassigned	741	25,05
Etiology	445	15,04
Association	282	9,53

Model definition and evaluation

The STs distribution within the axes has allowed the definition of about 35 slots in the model. For each term, morphemes are attributed to a slot. All the retained definitions of chapter XI could have been processed in the model. For example:

K22.6: **Gastroeosophageal laceration haemorrhage syndrome**, is processed as:

Pathology::*Macro*:**Laceration**
Pathology::*Sign*:**Haemorrhage**
Pathology::*Proc*:**Syndrome**
Location::*Organ*:**Gastro**
Location::*Organ*:**Oesophageal**

This code includes also a secondary term: **Mallory-Weiss Syndrome**, which gives:

Pathology::*Proc*:**Syndrome**

Proper noun:**Mallory**
Proper noun:**Weiss**

Discussion

The suggested semantic model allows us to represent a reasonable amount of the knowledge within ICD-10 term in a specific chapter. Considering the scalability issue, the model has been rapidly tested on some terms from other chapters with uneven results. There seems to be no particular problems for those sharing a common structure with chapter XI. Besides, for some other chapters, specific concepts, chromosomal alteration for example, could be modelled with additional slots. On the other hand, representing chapters relative to mental disorders, social problem or external causes of morbidity would imply drastic changes in the model.

Although French language is used in this work, the model seems to be appropriate for English and maybe other languages as well. As the model is based on morphemes, which are similar amongst medical terminologies, it may possibly be applied to other classifications.

However, the knowledge representation provided by this model is not entirely satisfying. The approach chosen for this work is to represent each ICD-10 terms separately, therefore neither include relationships between terms nor exclude and dagger-star relationships between codes are taken into account. Morphemes extraction is an automatic process but bringing some of them together and coupling them with STs is partially manual, therefore adding an important bias.

A four axes representation of terms has been motivated by the recurrent relations between a category and subjacent codes. However, such a representation leaves about a quarter of morphemes unassigned. Yet, most of these morphemes are related to negations or relative complement that could be also modelled. Though, even if the model could include morphemes relative to those concepts, a slot could not correctly model inherent information within expressions like "not elsewhere classified".

This work is based on semantic frames though it couldn't be considered as a strict application as many properties inherent to this kind of structure are left over. The model should rather be seen as a way to order the knowledge contained inside a term, with no reasoning capabilities.

Several medical NLP tools already exist (see, for a survey [14]) but automatic ICD-10 encoding from a patient discharge has not yet reached enough reliability to be presently used in hospitals [15]. This model, connected with the tool used for morpheme extraction, could be valuable for this kind of application but has not yet been implemented. So, even if significant results are obtained in modelling knowledge included in one chapter of ICD-10 and a larger part seems to be possibly modelled that way, evaluating the model consistency for a coding application still has to be done.

Conclusion

The increased demand in diagnoses encoding using ICD-10 call for reliable coding tool. The objective of this work is to show that a semantic model may be useful for this kind of applications by representing knowledge in an efficient way. However, the source for this model is a classification originally meant for epidemiological purposes and manual exploitation. Thus, its very structure state numerous problems which have to be overcome and a lot of work still has to be done in order to create a reliable coding tool using ICD-10.

Acknowledgements

Part of this work has been funded by the Swiss National Science Foundation (SNF 632-066041).

References

[1] Classification statistique internationale des maladies et des problèmes de santé connexes, dixième révision. Genève: Organisation Mondiale de la Santé, 1993.

[2] Lovis C, Griesser V, Michel PA, Rossier P, Borst F, Baud R, Scherrer JR. Codification des diagnostics et procédures: évaluation et implémentation d'une solution globale. In: Informatique et Santé, Springer-Verlag, ed. Vol. 8. PARIS, 1996:99-110.

[3] Michel PA, Lovis C, Baud R. LUCID: a semi-automated ICD-9 encoding system. Medinfo 1995;8 Pt 2:1656.

[4] Blanquet A, Zweigenbaum P. A Lexical Method for Assisted Extraction and Coding of ICD-10 Diagnoses from Free Text Patient Discharge Summaries. J Am Med Inform Assoc 1999; 6(suppl.).

[5] Landais P, Jais JP, Frutiger P. Sémantique des classifications et nomenclature. In: Informatique et Santé, Springer-Verlag ed. Vol. 2. PARIS, 1989; 1:211-22.

[6] Baud R, LovisC, Weber P, Griesser V. On the need of a reference ICD-10 Database. Proceedings MIE 2002.

[7] Unified Medical Language System Knowledge Sources: National Library of Medecine, 2002.

[8] Baud R, Lovis C, Rassinoux AM, Michel PA, Scherrer J-R. Automatic Extraction of Linguistic Knowledge from an International Classification. Medinfo 1998:581-5.

[9] Bouchet C, Bodenreider O, Kohler F. Integration of the analytical and alphabetical ICD10 in a coding help system. Proposal of a theoretical model for the ICD representation. Medinfo 1998;9 Pt 1:176-9.

[10] Petersson H, Nilsson G, Ahlfeldt H, Malmberg BG, Wigertz O. Semantic modeling of a traditional classification: results and implications. Medinfo 1998;9 Pt 1:613-7.

[11] Minsky M. A Framework for Representing Knowledge. In: The Psychology of Computer Vision. P. Winston, Ed., McGraw Hill, New York, 1975.

[12] The Systematised Nomenclature of Human and Veterinary Medecine: SNOMED International. Côte R, Rothwell D, Palotay J, Beckett R, Brochu L, Northfield, 1993.

[13] Ruch P, Wagner J, Bouillon P, Baud R, Rassinoux AM, Scherrer JR. MEDTAG: tag-like semantics for medical document indexing. Proc AMIA Symp 1999:137-41

[14] de Bruijn B, Martin J. Getting to the (c)ore of knowledge: mining biomedical literature. Int J Med Inf 2002;67(1-3):7-18.

[15] Franz P, Zaiss A, Schulz S, Hahn U, Klar R. Automated coding of diagnoses--three methods compared. Proc AMIA Symp 2000:250-4.