Medical Informatics Europe '99
P. Kokol et al. (Eds.)
IOS Press, 1999

Maintenance of self-consistency of coding tables by statistical analysis of word co-occurrences

György Surján, Gergely Héja
Haynal Univ. for Health Sciences P.O.B 112 Budapest Hungary H-1389

Abstract. The author presents a method for maintaining the internal consistency of coding tables. The method was tested on a table used for assisting the daily work of indexing clinical cases to International Classification of Diseases. 300 item were tested selected randomly form a corpus of 3082 clinical diagnoses. The method discovered potential consistency problems in 39 cases, out of which 10 were false positive.

1. Introduction

Computer systems assisting the work of coding diseases usually contain some sort of coding tables, consisting of a list of clinical terms and the corresponding codes. Such tables are growing in size by time to time. The updating process usually leads to loss of the internal consistency. The correct semantic representation of all concepts would be necessary to solve the problem. There are promising ongoing works in this field. (Like in the GALEN-IN-USE project [1].) However it will be a long way, when the results of these projects can be used in the everyday practice. Therefore, it seems to be useful to search for more simple methods, which do not solve the problem in an exact way, but can be still helpful.

2. The method of word co- occurrence measurements

The word co-occurrence analysis is commonly used method in computational linguistics [2]. This usually based on so called *n*-grams, that is a window of *n* consecutive words. Instead of n-grams, our approach is sentence based. The code values are included in the text as if they where regular words. For example, the last item in the left table in Figure 1 can be considered as a sentence, consisting of three words: *aortic, vitium* and *I069.* In order to be able to discriminate from normal words, all code values are placed within $ signs: *$I069$ aortic vitium.* Now we can represent our table by a Sentence × Word matrix, in which the sentences form the columns, and the words appearing anywhere in the whole corpus form the rows. The value of the matrix cells is *1* when the corresponding word appears in the corresponding sentence, and *0* if not. (Figure 1.)

This vector space model is used in *latent semantic indexing* [3]. In this approach the cosine of the angle between the vector of two documents (sentences) represents their semantic similarity. We consider the Sentence × Word matrix as a set of word-vectors in the space of sentences. Then the cosine of the angle between the word vectors represents the

"co-occurring tendency" of the corresponding word pairs. The formula that defines the cosine values is as follows:

$$\frac{\mathbf{w}_i \circ \mathbf{w}_j}{\|\mathbf{w}_i\| * \|\mathbf{w}_j\|}$$

where $\mathbf{w}_i$ and $\mathbf{w}_j$ the two Word vectors, ° represent their scalar product, and ‖ ‖ symbol means the Eucledian norm of the vector. The resulting Word × Word matrix is necessarily symmetric. If the presence of one of the words is obvious, another measure can be used, according to the following formula:

$$\frac{\mathbf{w}_i \circ \mathbf{w}_j}{\|\mathbf{w}_j\|^2}$$

Table 1 shows the result of this calculation for our demonstrating example. Clearly, this matrix is asymmetric. The word '*aorta*' appears whenever '*vitum*' appears, but the reverse is not true. (The 'given word' is represented by the columns.)

the original table → conversion into "sentences" → Sentence × Word matrix

ICD code	language expression
I.700	aorta atherosclerosis
I.069	aorta vitium

S1	$I700$ aorta atherosclerosis
S2	$I069$ aorta vitium

	S1	S2
$I700$	1	0
$I069$	0	1
aorta	1	1
atherosclerosis	0	1
vitium	1	0

Figure 1 The representation of tables by Sentence × Word matrix

A sufficiently large corpus of sentences can be considered as training set representing some kind of knowledge about word co-occurrences. Any expression consisting of the same set of words can be considered as testing sample. Now we define the 'environment' of a sentence. This is a set of all words of the training set, which have co-occurrence likelihood with any of the words of the analysed sentence above a given threshold. If any word (or code) is missing by mistake, it likely will appear in the environment. For all words in the environment the co-occurrence tendency to all words of the tested expression can be calculated. These values form the *environment matrix* (Figure 2.)

The conditional co-occurrence likelihood values

	$I700$	$I069$	aorta	atherosclerosis	vitium
$I700$	1	0	0.5	1	0
$I069$	0	1	0.5	0	1
aorta	1	1	1	1	1
atherosclerosis	1	0	0.5	1	0
vitium	0	1	0.5	0	1

Table 1

The geometric mean of the co-ordinates of the word vectors of the environment matrix can serve as a measure for how strongly belongs the corresponding word to the analysed sentence.

3. Application of co-occurrence measurements to the problem

The table, which is subject of the consistency analysis, has be used as training set. Then, all codes should be removed from all sentences, and the rest serves as testing set. Now we have two assumptions:

Assumption 1:The original code should appear in the environment, with non-zero co-occurrence likelihood to all words of the sentence. This is self-evident.

Assumption 2: When the training set contains ambiguity, then apart from the original code, some others will also appear in the environment.

Words	$K37H0$	Appendicitis	geometric mean
$K3500$	0.	0.25	0.
$K3590$	0.	0.25	0.
acuta	0.	0.5	0.
Consec.	0.5	0.25	0.3536
ulcero-phlegmonosa	0.	0.25	0.

Figure 2. The environment matrix of the expression "$K37H0$ Appendicitis".

The second assumption is the subject of our study. The experiment was performed by means of a special software was developed by the authors to analyse word co-occurrences. Technical details are not discussed here.

A table containing 3082 clinical diagnosis expressions was used as training set. The diagnoses were taken from discharge reports collected from five different departments of our hospital. They where assigned to International Classification of Diseases (ICD-10) by a physician. The 'sentences' contain the words of the diagnostic expression with one and only one ICD code. The test sample was 300 diagnosis randomly selected from the training set. The candidates for error according to assumption 2 were detected, and analysed in details. E.g. in the environment of "Art. scler. cerebri" (cerebral atherosclerosis) two codes were found: the original I6720 and I2510 with higher co-occurrence tendency. The training set was searched for occurrences of $I2510$. The expression shown in the rectangle in Figure 3 is a combination of two different diseases in a single diagnostic statement: 'arterioscleoris coronarium (I2510) and 'arteriosclerosis cerebri' (I6720). Such compound diagnoses should be split up to avoid ambiguity. (The essential difference between *diagnosis* and *disease* is described by us in a previous paper [4].)

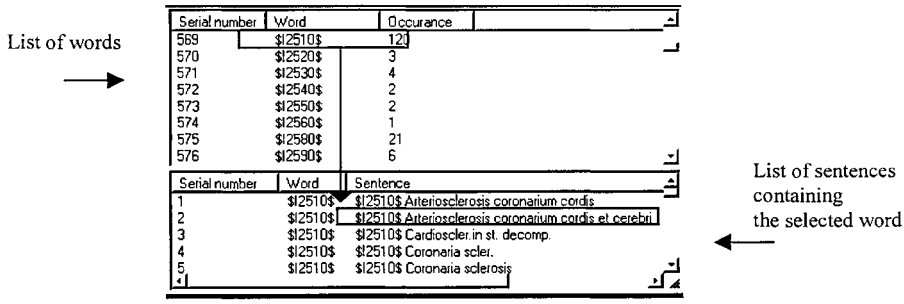

Figure 3 Occurrences of the code I2510

The different reasons why non-original codes emerge in the environment can be categorised in four main groups: 1) error or inconsistency of the training set; 2) inconsistency of the ICD; 3) false positive result. Group 3) can be divided into two subgroups: 3a) the emerging codes are alternatives of the original; 3b) true false positive result. A case belongs to 3a)

only if all the emerging codes have lower co-occurrence tendency than the original and are semantic relatives of it.

4. Results

A) The original code always appeared in the environment. This is in accordance with Assumption 1, and proves that the software does not make serious errors.
B) No other codes appeared in the environment in 87% (261 cases). In these cases the method was not able to detect consistency problems in the coding table.
C) Other codes appeared with lower affinity than the original in 5.6% (17 cases)
D) Other codes appeared with equal or higher affinity than the original in 7.33% (22 cases)

The cases C) and D) were analysed manually and the reasons were categorised using the above described groups. The result is presented in Table 2. Note that one case might belong to more than one category. The most frequent error type was the ambiguous indexing of the same concept into different ICD category.

Causes of appearance of non-original codes in the environment

co-occurrence tendency of the other codes:	Number of cases	1) Error in reference sample		2) inconsistency of ICD		3a) alternative codes		3b) false positive cases	
C) lower	17	10	45.45%	3	13.64%	7	31.82%	0	0%
D) equal or higher	22	14	63.64%	7	31.82%	-	-	10	45.45%
Total	39	24	61.54%	10	25.64%	7	17.95%	10	25.64%

Table 2

5. Discussion

The presented result convinces us that it is possible to reduce but not eliminate inconsistency of the coding tables by this method. Considering the computational requirements, the cost benefit is questionable, but the same approach can be used for other purposes. We were able to detect inconsistencies in the ICD as well. The most typical problem was here the lack of generic concepts. Our approach with some modifications can be used also as a tool for assisting the work of encoders.

6. Acknowledgements

Special thanks to Prof. Arie Hasman for assisting this work and commenting the draft of this paper. The recent research was supported by the grand of the Hungarian Ministry of Health. (Grant No.: 155/98)

References

[1] Rector AL Compositional Models of Medical Concept: Towards Re-Usable Application - Independent Medical Terminologies in Knowledge and Decisions in Health Telematics, (Eds) P Barahona and J P Christensen, IOS Press 1994. pp 109 - 114
[2] J Allen Natural Language understanding Addison Wesley 1995
[3] Chute C G, Yang Y An overview of Statistical Methods for the Classification and Retrieval of Patient Events Methods of Information in Medicine (1995); 34:104-110
[4] Surján G Questions on validity of ICD coded diagnoses International Journal of Medical Informatics54 (1999) 77-95