Published in final edited form as:
J Biomed Inform. 2012 February ; 45(1): 61–70. doi:10.1016/j.jbi.2011.08.021.

Overcoming an Obstacle in Expanding a UMLS Semantic Type Extent

Yan Chen, PhD[1], **Huanying Gu, PhD**[2], **Yehoshua Perl, PhD**[3], and **James Geller, PhD**[3]
[1]CIS Department, Borough of Manhattan Community College, CUNY

[2]Department of Computer Science, New York Institute of Technology

[3]Department of Computer Science, New Jersey Institute of Technology

Abstract

This paper strives to overcome a major problem encountered by a previous expansion methodology for discovering concepts highly likely to be missing a specific semantic type assignment in the UMLS. This methodology is the basis for an algorithm that presents the discovered concepts to a human auditor for review and possible correction. We analyzed the problem of the previous expansion methodology and discovered that it was due to an obstacle constituted by one or more concepts assigned the UMLS Semantic Network semantic type **Classification**. A new methodology was designed that bypasses such an obstacle without a combinatorial explosion in the number of concepts presented to the human auditor for review. The new *expansion methodology with obstacle avoidance* was tested with the semantic type **Experimental Model of Disease** and found over 500 concepts missed by the previous methodology that are in need of this semantic type assignment. Furthermore, other semantic types suffering from the same major problem were discovered, indicating that the methodology is of more general applicability. The algorithmic discovery of concepts that are likely missing a semantic type assignment is possible even in the face of obstacles, without an explosion in the number of processed concepts.

Keywords

UMLS; semantic type assignment; auditing; group auditing; neighborhood auditing; refined semantic type

1 Introduction

The Unified Medical Language System (ULMS) [3, 4, 15, 16] is a very large and complex terminological system for biomedicine. It consists of two layers, the Metathesaurus (META) [24, 25], which is a repository of concepts, and the Semantic Network (SN) [17, 18], which is a compact abstraction network consisting of a small number (133) of broad categories called semantic types (STs). The connection between the layers is implemented by assigning each concept one or more semantic types.

The assignments of STs to concepts play a major role in the integration of new terminologies into the UMLS. Due to the extensive size and complexity of the UMLS, errors are inevitable. Auditing is therefore essential to ensure the quality of the UMLS. The ST

Reprint Requests & All Communications: Yan Chen CIS Department Borough of Manhattan Community College, CUNY 199 Chamber Street, New York, NY 10007 Tel: (212) 220-8384 Fax: (212) 220-1287 ychen@bmcc.cuny.edu.

assignments were proven instrumental in auditing the UMLS for various errors [7, 11, 12, 13, 14]. ST assignment errors, including incorrect and missing ST assignments, were discovered [6, 8, 13, 14, 23]. Redundancy, circularity, omissions and other problems in hierarchical relationships were located [1, 2, 7, 20]. Classification errors were found [9, 10, 13]. Tools such as the Neighborhood Auditing Tool (NAT) [19] have been developed to facilitate auditing. For an extensive review of auditing of terminologies in general and the UMLS in particular, see [26].

In a study of uses of the UMLS [5], users expressed that incorrect and missing semantic type assignments are errors of greatest concern. Certain structural configurations indicate concepts with a high likelihood of incorrect ST assignments [6, 13, 14]. However, for the problem of exposing concepts with missing ST assignments there are no structural indicators.

The difficulty of exposing missing ST assignments was demonstrated by the findings of Chen et al. [8], where about thousand concepts of the UMLS that had been correctly assigned **Neoplastic Process**[1](**NP**)[2] were missing the assignment of the second ST **Experimental Model of Disease** (**EMD)**. Those concepts were mainly experimental cancers in mice. They were integrated into the UMLS from the National Cancer Institute thesaurus (NCIt), where they are in the Experimental Organism Diagnoses (EOD) hierarchy. The NCIt maintains its own ST assignments. According to Mougin and Bodenreider [21], these assignments differ from the UMLS ST assignments for some concepts and were proven more accurate. However, the **EMD** assignments were missing for those approximately thousand concepts in the NCIt as well.

In previous research we corrected some **EMD** assignments, but did not detect the concepts missing **EMD** [12]. Furthermore, in 2004, our team audited the Experimental Organism Diagnoses hierarchy of the NCIt for missing relationships and still did not detect the missing ST assignments. The difficulty of detecting concepts with missing ST assignments stems from the lack of a suspicious configuration which indicates their absence, in contrast to the existence of structural indicators for detecting incorrect ST assignments. Without such an indicator, an auditor receives no guidance where to search for missing ST assignments. Searching in an arbitrarily selected part of the UMLS is likely to offer a low yield for an extensive effort.

In the work of Chen et al. [8] we presented a methodology for finding concepts with missing ST assignments. It is based on the assumption that a concept that is in the neighborhood of other concepts that are already assigned a specific ST, but does not have this assignment, very likely *should* have this ST assigned. Furthermore, this process was dynamic; once a concept had been assigned the additional ST, its neighbors were also reviewed [8]. For more details, see the Background Section.

In spite of our success in algorithmically discovering many concepts missing **EMD** assignments, confirmed by human auditors [8], not all concepts in the Experimental Organism Diagnoses hierarchy of NCIt missing this assignment were discovered. There are hundreds of experimental diseases (mainly cancers of different kinds) in rats which should be assigned **EMD** and are currently assigned **NP** for cancers or **Disease or Syndrome** (**DS**)

[1]Semantic types are written in bold, while concept names are in italics.

[2]The following abbreviations are used in the paper: **CL** (Classification), **DS** (Disease or Syndrome), **EM/OA** (Expansion methodology with obstacle avoidance), **EMD** (Experimental Model of Disease), **EOD** (Experimental Organism Diagnoses), **HPS** (Hazardous or Poisonous Substance), **NAT** (Neighborhood Auditing Tool), **NCIt** (National Cancer Institute thesaurus), **NP** (Neoplastic Process), **OC** (Organic Chemical), **RST** (Refined Semantic Type), **SN** (Semantic Network), **ST** (Semantic Type), **SV** (Secondary enVelope), **UMLS** (Unified Medical Language System), and **XV** (auXiliary envelope).

for non-cancer experimental diseases. Of course, once this fact has been exposed, one could screen this hierarchy and correct the ST assignments of these concepts but we would like a methodology for the detection of such cases. In this paper, we are presenting such a methodology for discovering missing ST assignments. When we analyzed what prevented our previous methodology from reaching the missed concepts, we found that an "obstacle" was separating the discovered concepts from those which were not discovered. In this paper, we present a methodology that bypasses such an obstacle and reaches the concepts behind it that are missing the correct ST assignments. The results of applying this methodology for **EMD** are reported. This methodology is applicable to other STs to discover more missing ST assignments.

2 Background

2.1 The Refined Semantic Network for the UMLS

In the UMLS, each concept is assigned at least one semantic type. The set of all concepts that are assigned the same ST is called its *extent*. However, the concepts in the extent of an ST are not necessarily assigned only that ST. For example, the concepts *Arthritis, Experimental* and *Experimental Hepatoma* are in the extent of **EMD.** However, *Experimental Hepatoma* is also assigned **Neoplastic Process**. Therefore, these two concepts do not share the same semantics (expressed by the ST assignment) even though they are both in the extent of **EMD**. Hence, the extent of **EMD** is not semantically uniform.

To achieve semantically uniform sets of concepts, each extent needs to be partitioned into subsets to reflect a refinement of this ST. We proposed the Refined Semantic Network for the UMLS, consisting of Refined Semantic Types (RSTs) for this purpose [11, 12], consisting of Refined Semantic Types (RSTs). Each RST is either a "Pure Semantic Type" or an "Intersection Semantic Type." Each Pure Semantic Type corresponds to one ST from the SN and is assigned to concepts that were *only* assigned this one ST in the UMLS. All concepts with multiple ST assignments are removed from the extent of the Pure Semantic Type. An Intersection Semantic Type is a combination of two or more STs from the SN and its extent contains concepts assigned exactly such a combination of STs. Hence, in contrast to the extents of the original STs, the extent of each RST contains the concepts that are *only* assigned this RST and have the semantics expressed by it.

Our previous auditing methodology, reported by Chen et al., expands the extent of an ST by separately expanding each of its RSTs [8]. This expansion process identifies any neighboring concepts that have the same semantics as the concepts in the RST's extent and inserts them into the extent. The semantic uniformity of RSTs' extents makes human auditing of the concepts in those extents more effective and efficient.

2.2 Methodology for Expanding the Extent of a Semantic Type

In the work of Chen et al. [8], a two-part methodology was introduced for aiding an auditor in discovering missing ST assignments, by narrowing down the set of concepts presented to him. The auditing focused on a neighborhood surrounding the extent of an RST[3] $\mathbf{T^R}$ (E($\mathbf{T^R}$)) called an *envelope* (denoted as V($\mathbf{T^R}$)), consisting of neighbors, i.e. parents and children of the concepts in the extent which are themselves not in the extent. All concepts in an envelope are audited by a human expert. If a concept with a missing ST assignment is identified then it is corrected and the neighbors of this concept are inserted into the next envelope.

[3] $\mathbf{T^R}$ is the **R**efined (in this case "pure") semantic type of the semantic type **T.**

Part 1 of the auditing methodology can be depicted as expanding outward from an extent in a series of concentric circles, as shown for $\mathbf{EMD^R}$ (Figure 1). For example, *Arthritis, Animal Model* and *diencephalic hyperactivity* reside in V($\mathbf{EMD^R}$). An auditor finds that *Animal Model* is lacking the assignment of $\mathbf{EMD^R}$. Thus, its parents, Animal Study, in vivo Model, Investigative Techniques and *Study models* and its children, *Dorsal Skin Fold Window Chamber Model* and *Olfactory Learning*, not already in E($\mathbf{EMD^R}$) or V($\mathbf{EMD^R}$), are included in the second-level envelope V^2($\mathbf{EMD^R}$) and await auditing after the processing of V($\mathbf{EMD^R}$) has been completed. If any concepts in V^2($\mathbf{EMD^R}$) are later found to be missing the assignment of $\mathbf{EMD^R}$, then their parents and children not already in E($\mathbf{EMD^R}$), V($\mathbf{EMD^R}$), or V^2($\mathbf{EMD^R}$) will be entered into the third-level envelope V^3($\mathbf{EMD^R}$) that is processed after V^2($\mathbf{EMD^R}$). This process continues until the next envelope remains empty. Due to the auditing process, the concepts that in broken-line boxes in Figure 1 are assigned $\mathbf{EMD^R}$.

This methodology might lead to the assignment of an RST to a concept that is quite far from the concepts in the original extent of this RST. The condition for a concept *c* to be assigned $\mathbf{T^R}$ is that there exists a path of concepts connected by parent or child relationships from a concept *s*, originally assigned $\mathbf{T^R}$, all the way to the concept *c*, such that each intermediate concept on this path is also assigned $\mathbf{T^R}$. The expansion in a sequence of concentric envelopes implements the expansion process in a stepwise manner. Hence a “long distance” expansion is achieved via repeated local expansion steps.

The described process is efficient, since it does not expand in every direction for the longest possible distance. The stepwise expansion happens only for concepts where an ST assignment was made in the previous step. Hence, even if an expansion proceeds along a path of, say, ten concepts, the actual processing done is proportional only to the number of concepts, that were assigned the new RST and their parents and children, but not for all concepts within a distance of ten from the concept originally assigned the RST.

Part 2: As explained in Section 2.1, the extent of a semantic type **T** consists of disjoint subsets of concepts, such that there exists one subset for each RST generated from **T.** While reviewing the envelope of another RST of **T**, say, $\mathbf{T^{R2}}$, the auditor might realize that some of the concepts in the envelope of $\mathbf{T^{R2}}$ should be assigned $\mathbf{T^R}$ instead. These concepts are inserted into the *auxiliary extent* of $\mathbf{T^R}$, denoted AUX($\mathbf{T^R}$). Once Part 1 of the methodology on all other RSTs of **T** has been completed, the auditor processes the auxiliary extent AUX($\mathbf{T^R}$). This set is processed in a manner analogous to E($\mathbf{T^R}$). For instance, the concept *Mouse Mammary Gland Disorder* in V^5(**EMD∩NP**) is mis-assigned **Disease or Syndrome**. This concept should be assigned $\mathbf{EMD^R}$ instead and is therefore inserted into AUX($\mathbf{EMD^R}$). The consecutive envelopes of AUX($\mathbf{EMD^R}$) are constructed in the same manner as in Part 1. Figure 2 illustrates the processing of AUX($\mathbf{EMD^R}$). Only some concepts of AUX($\mathbf{EMD^R}$), V(AUX($\mathbf{EMD^R}$)), and V^2(AUX($\mathbf{EMD^R}$)) are shown.

3 Methods

3.1 A Major Problem with the Previous Expansion Methodology

Despite of the success of the expansion methodology of Chen et al. for **EMD**, it failed to discover all the concepts of the NCIt hierarchy “Experimental Organism Diagnoses,” which deal with experimental cancer diagnoses [8]. To understand the reason for this major problem, see Figure 3, showing a partial indented hierarchy, as shown by the NCIt browser [22] into which we have added UMLS ST assignments between {}. Due to their general nature, the concepts *Experimental Organism Diagnosis*, which is the root of this hierarchy, its child Rat Histopathology Diagnoses for Proliferative Changes, and its grandchild *Rat Neoplams by Morphology* are assigned **Classification,** which is defined as “A term or

system of terms denoting an arrangement by class or category." The latter concept is the parent of nine concepts, which are assigned **NP** and should be assigned **EMD∩NP.** These three concepts, correctly assigned **Classification,** constitute an obstacle which stops the dynamic expansion process starting at *Rous Sarcoma*, assigned **EMD∩NP,** from reaching these nine concepts Thus, the described expansion methodology cannot go beyond this obstacle, and as a result concepts are missed [8].

However, by modifying the expansion process to go from *Rous Sarcoma* beyond its parent *Experimental Organism Diagnosis* and the child and grandchild of the latter, Rat Histopathology Diagnoses for Proliferative Changes, and *Rat Neoplams by Morphology*, the process will reach the subhierarchy of hundreds of concepts rooted in *Rat Neoplasm by Morphology*, which are neoplasms in rats, and their ST assignments will be corrected to **EMD∩NP**. Note that two of the children of *Rat Neoplasm by Morphology*, *Rat Carcinoma* and *Rat Adenoma* have further children (see Figure 3). The question is how the previous expansion process should be modified to bypass such an obstacle.

3.2 Expansion Methodology with Obstacle Avoidance

In order for the expansion process to bypass an obstacle, we need to formulate it such that concepts that are assigned **Classification** do not interrupt the expansion, while leaving their semantic type assignments unchanged. In other words, the expansion process should bypass the obstacle, without affecting the ST assignments of the concepts of the obstacle. We call this process *expansion methodology with obstacle avoidance (EM/OA)*. An obstacle may consist of one or several concepts assigned **Classification.**

As a solution, we introduce a secondary envelope SV(**T**) and an auxiliary envelope XV(**T**) for this purpose. The secondary envelope will contain concepts constituting an obstacle. Since an obstacle may consist of a path of several nodes, we will represent the number of its hierarchical levels ("the width of the obstacle") using SV^1(**T**), SV^2(**T**), etc. parents and children of concepts in the secondary envelope that are not already in E(**T**), V(**T**) or XV(**T**) are inserted into one auxiliary envelope XV(**T**), without levels.

A parameter *p* defines how many SV^i(**T**) are allowed. The value for *p* will be determined experimentally for different situations. For each concept in SV^i(**T**), $i \leq p$, its parents and children are inserted into the auxiliary envelope XV(**T**). We let the auditor review them for potential corrections. If such a parent or child should be corrected to the RST **T**, then it would be entered into the extent of **T**. If this concept is correctly assigned **Classification**, then it will be entered into SV^{i+1}(**T**) unless $i=p$. By the definition of **Classification**, it is the only ST which has the potential to categorize a high level concept, not by its meaning but by its role, representing a group of concepts in the terminology that have a joint meaning.

By limiting *p* to a small number, we limit the number of concepts of XV that are audited. In practice, *p* will be chosen experimentally to be large enough to bypass obstacles but small enough to avoid large scale auditing. The best value for *p* might have to be determined by an iterative process.

Also, by constraining the obstacles to concepts assigned **Classification**, we further limit the auditing effort. If **Classification** assignments are incorrect the auditor will replace them and the corresponding concepts will cease to be obstacles. If a concept in XV(**T**) is neither found to be assigned **Classification** nor needs to be corrected to **T**, it is discarded from the auditing process. Once the auxiliary envelope XV(**T**) is empty, the auditing algorithm stops.

We will now use the example of **EMD∩NP** to show how the EM/OA will reach Rat neoplasm concepts in the Experimental Organism Diagnosis hierarchy of NCIt. We illustrate

how the methodology bypasses the obstacle of concepts assigned **Classification** and reaches the Rat experimental cancer disease concepts that were not reached by the methodology of Chen [8] (Figure 3).

As demonstrated in Figure 4, the concept *Rous Sarcoma*, assigned **EMD∩NP**, is the starting point of the EM/OA, and we will use *p=3* to limit the number of secondary envelopes.[4] Its parent, *Experimental Organism Diagnosis*, correctly assigned **Classification**, is inserted into SV^1(**EMD∩NP**). Then the other children of this concept, *Experimental Allergic Encephalomyelitis* (**EMD**), *Mouse Pathologic Diagnoses* (**DS**) and Rat Histopathology Diagnosis for Proliferative Changes (**Classification**), are inserted into XV(**EMD∩NP**). From those, the concept Rat Histopathology Diagnosis for Proliferative Changes is the only one inserted into SV^2(**EMD∩NP**), because it is assigned **Classification.** Figure 4 displays Rat Histopathology Diagnosis for Proliferative Changes after it was moved from XV(**EMD∩NP**) to SV^2(**EMD∩NP**).

Next, the children of Rat Histopathology Diagnosis for Proliferative Changes (Figure 3) are inserted into XV(**EMD∩NP**). From those children, three concepts in bold boxes in XV(**EMD∩NP**) that are assigned **NP**, Rat Unclassifiable Benign Tumor, Rat Unclassifiable Malignant Tumor and *Tissue Autolysed Diagnosis Not Possible* are reassigned **EMD∩NP**. None of those three concepts have children, so the downward expansion stops.

Three other children, in bold, rounded corner boxes, are assigned **Classification**, Rat Neoplasms by Location, Rat Neoplasms by Morphology and *Rat Proliferative Change by Location*. Figure 4 shows the status before they are inserted into SV^3(**EMD∩NP**). In Figure 5, where the three **Classification**-assigned concepts, shown as rounded corner boxes, appear in SV^3(**EMD∩NP**), the continuation of applying the EM/OA algorithm is demonstrated. When these concepts are processed, their children are inserted into XV(**EMD∩NP**). When, in turn the children of one of them, *Rat Neoplasms by Morphology* are audited, they are reassigned **EMD∩NP**. The continuation of the expansion process from two of these concepts *Rat Adenoma* and *Rat Carcinoma* which have children (as seen in Figure 3) is straightforward, since no more obstacles are encountered.

4 Results

4.1 Expansion of E(EMD∩NP) with Obstacle Avoidance

Table 1 describes the expansion of E(**EMD∩NP**). Rows 1 to 12 are taken from Chen et al., Table 3 [8]. Stages 1-3 summarize the obstacle avoidance process. The rows below them describe the concepts audited after obstacle avoidance, using the same methodology as before [8].

4.2 Expansion of E(EMD) with Obstacle Avoidance

A similar auditing process happens for the mouse non-cancer experimental diagnosis concepts, where **DS** is replaced by **EMD**. These concepts were previously not reassigned **EMD** due to an obstacle. This process is initiated by the concept *Experimental Allergic Encephalomyelitis*, the child of *Experimental Organism Diagnosis*, assigned **EMD**, following the auditing methodology (Part 1 and 2), with the 165 concepts in E(**EMD**).

During the obstacle avoidance process, concepts with missing **EMD** assignments were identified both in XV(**EMD**) and V(**EMD**) at Stage 3. In Table 2, there are two additional columns for the number of concepts added to SV(**EMD**) and XV(**EMD**) at each auditing

[4] SV(**EMD∩NP**) = SV^1(**EMD∩NP**)

stage. The Error Rate = (#Concepts added to E(**EMD**)) / (#Concepts in XV(**EMD**) + #Concepts in V(**EMD**)). Stages 1 to 3 show the obstacle avoidance process. V^1 and V^2 describe the auditing after obstacle avoidance, following the methodology of [8].

4.3 Restarting the Expansion of E(EMD∩NP) Due to Audit of E(EMD)

As shown in Table 1, at the end of the expansion process, there were 1083 concepts in the extent of E(**EMD∩NP**). In the process of expanding E(**EMD**), ten concepts, e.g. *Mouse Neoplasm*, were added to E(**EMD∩NP**) (Figure 3), resulting in 1093 concepts in E(**EMD∩NP**). These ten concepts serve as starting points of a second round of expansion of E(**EMD∩NP**), according to Part 2 of our methodology [8]. Table 3 shows the results of expanding these ten concepts, resulting in an extent E(**EMD∩NP**) of 1397 concepts. The list of concepts requiring changes in their ST assignments were submitted to the NLM and NCI.

5 Discussion

This paper overcomes a major problem encountered in the process of expanding the extent of an RST **T** [8] during auditing. That process used the envelope of this extent as a subset of concepts of high likelihood to require the **T** assignment. The expansion process may be blocked by an obstacle consisting of concepts assigned **Classification**. In this paper, we present a methodology for bypassing such an obstacle, so that the **T** assignment is propagated behind the obstacle, without changing the ST assignment of concepts in the obstacle itself. For this, we used a parameter *p*, which controls how many levels of concepts in the obstacle may be bypassed.

Classification is an unusual ST that does not represent the semantics of a concept, as other STs do, but the role of a concept to classify its descendants. This makes it possible to have a subhierarchy of META concepts, being rooted in a generic concept describing them, by assigning the semantic type **Classification** to the generic concept. Furthermore, its children and sometimes even grandchildren may also be generic concepts, describing smaller subsets of this META subhierarchy. All these generic concepts may legitimately be assigned **Classification**. This observation allows singling out **Classification** as the source of obstacles to expanding the assignment of an ST from one branch of a UMLS concept hierarchy to another branch, and is the basis for the assumption that the number of levels with such concepts is typically small. In Figure 1, the concept *Rat Neoplasm by Morphology*, assigned **Classification,** has a parent and grandparents assigned **Classification**, but no children assigned **Classification**. To bypass such an obstacle of three levels, a parameter *p*=3 is needed.

By using a small *p*, we limit the scope of the human auditing effort, since the larger *p* is, the more concepts need to be reviewed by the auditors. We compared the impact of two values of *p* for the example of Section 3.2. For *p*=2 and *p*=3, the numbers of concepts entering SV were 2 and 6, respectively, causing their respective 13 and 32 neighbors to enter XV, of which 3 and 12 concepts, were reassigned **EMD∩NP.** This small example illustrates the growth in the number of audited concepts with the increase of the *p* value. No advantage of using *p*=4 exists, since no additional descendants are assigned **Classification.** However, with *p*=2, the process did not succeed in bypassing the obstacle. Hence, we searched experimentally for the lowest *p* value for which the obstacle is bypassed.

Interestingly, it is also possible to bypass this obstacle with *p*=2, by two consecutive passes. In the first phase, the three concepts (in bold boxes in Figure 4) are assigned **EMD∩NP**. Although those concepts have no children, one of them can be used as a starting point for the second pass of the obstacle avoidance process, which bypasses the obstacle to reach all **NP**-assigned children of *Rat Neoplasm by Morphology* assigned **Classification**. This

enables the expansion of the **EMD∩NP** assignment from those children similar to what was shown Section 2.2.

The total number of concepts added or reassigned **EMD** according to Tables 1–3 is 554. The ten children of *Rat Proliferative Change by Location* in Figure 3 were reassigned **EMD** in Table 2. Alternatively, they may be reassigned **EMD** while bypassing the obstacle. In such a case they would be the basis for expansion according to Part 2 of the methodology [8], corresponding to the results in Section 4.3.

By the rules of the UMLS, an assignment of an ST **A** to a concept, which is also assigned a descendant of **A**, is redundant and forbidden. In recent years, the NLM has eliminated such redundant semantic type assignments from META. According to Dr. S. Srinivasan[5] from the NLM, the NLM is using a program to test each new UMLS release for redundant semantic type assignments. If any are found, they are eliminated. When our expansion methodology is applied, it avoids redundant semantic type assignments. Thus, the IS-A relationships in the Semantic Network do not have the negative effect of creating redundant ST assignments in our methodology. For example, our methodology would not assign both **Neoplastic Process** and its parent **Disease or Syndrome** to the same concept. This paper focused on **EMD**. More research is needed to explore other parts of the UMLS where this methodology is applicable. For example, consider *Organic Carcinogen* assigned **Classification**, whose parent, *Carcinogens*, is assigned **Hazardous or Poisonous Substance** (**HPS**). *Organic Carcinogen* has 29 children, seven of which are assigned **Classification**. Many other children are assigned **HPS** and **Organic Chemical** (**OC**). Two are assigned only **OC**. A similar picture appears for the children of the seven children assigned **Classification**. But a few of their children are again assigned **Classification**. We found an obstacle consisting of a chain of four concepts, namely Organic Carcinogen, Organo Nitrogen Carcinogen, Nitro Compound Carcinogen, and *Nitroarene Carcinogen*. The latter has five children assigned **HPS∩OC**. There are several such chains of length three. Two children of *Organic Carcinogen* assigned just **OC**, *Acetaldehyde* and *Vinyl Carbamate*, should also be assigned **HPS** as their grandparent, by their definition. This provides an example where the obstacle avoidance methodology would have spread **HPS∩OC** from the proper children of *Organic Carcinogen* to those two children missing **HPS**, bypassing the obstacle consisting of *Organic Carcinogen* only.

Another example is **OC**-assigned *Naphthalene*, used to manufacture moth balls. It follows a hierarchical chain of three ancestors assigned **Classification**. *Naphthalene* should be assigned **HPS** as its siblings, bypassing its **Classification**-assigned ancestors. These examples show STs other than **EMD** needing expansion, where concepts assigned **Classification** constitute an obstacle that our methodology bypasses. Further investigation of the 1664 concepts in the extent of **Classification** is likely to expose more potential ST targets for our expansion methodology.

It is not clear how to estimate how many concepts are blocked from receiving the correct ST assignments due to the obstacle concepts that are assigned **Classification**. For example, it is even difficult to estimate how many descendants of *Carcinogens* are missing an assignment of **HPS**, as illustrated in the examples above. There are 6054 descendant concepts of the concept *Carcinogens*, of which only 426 are assigned **HPS** and 26 are assigned **Classification**. Some concepts in the META subhierarchy of *Carcinogens*, like those representing chemicals, drugs or kinds of food are (correctly) not assigned **HPS**. Many such concepts are found, e.g., among the descendants of *Carcinogenic Mixture*, assigned **Classification**, which is a child of *Carcinogens*. For example, *Carcinogenic Mixture* has

[5]personal communication

children *Alcoholic Beverages* assigned **Food** and *Coal Tar* assigned **OC** and **Pharmacologic Substance**, each having many descendants.

However, it seems that many concepts in the META subhierarchy of *Carcinogens* are missing the **HPS** assignment. (We note that not all child-of relationships in the META are IS-A relationships; there are several other options.) To find out whether this is the case, one has to first apply the expansion methodology of [8] for the **HPS** ST. This methodology will require manual review by a domain expert. Only then will it be possible to see how many concepts, which do not have an assignment **Pharmacologic Substance**, or **Clinical Drug**, or **Food,** are still without the **HPS** assignment and potentially are missing it.

Using the list of 26 concepts assigned **Classification**, which are broad categories such as *Organic Carcinogen* and *Carcinogenic Hydrocarbon*, one can then apply the current methodology to overcome obstacles constituted by these 26 **Classification**-assigned concepts. It is difficult to estimate the number of concepts for which the **HPS** assignment will be added as a result of such a process, which would require substantial time of a domain expert. However, it seems that such a process will correct the ST assignments of a significant portion of the META subhierarchy of *Carcinogens.* We note again that the effort of domain experts required is limited to concepts that are corrected, and their neighbors. Hence, the yield of domain expert work, using our methodology, measured as the ratio of erroneous concepts to reviewed concepts, is expected to be high, as, for example, it was reported for the semantic type **EMD**.

Due to the unusual nature of **Classification**, there is inconsistency regarding its use. For example, the concept *Mouse Pathologic Diagnoses* and two of its children, are assigned **DS** (the third is assigned **NP**) (Figure 3). We note that corresponding concepts for Rats are assigned **Classification**. Furthermore, in the NCI thesaurus, which has its own independent ST assignment, those concepts are assigned **Classification**. On the other hand, *Carcinogens*, assigned **HPS,** is a broad category. Like its 26 descendants that are assigned **Classification**, it should have been assigned **Classification** to maintain the consistency of semantic type assignments.

In this research we did not try to correct **Classification** assignments, but limited ourselves to demonstrating the effectiveness of the obstacle avoidance methodology. Further research into the extent of **Classification** is likely to expose erroneous as well as missing **Classification** assignments.

The only other ST that would potentially constitute an obstacle for our expansion process, is **Conceptual Entity (CE)**, a child of **Entity**. The definition of this ST is “A broad type for the grouping of abstract entities or concepts.” As for **Classification**, the emphasis is on grouping, but the difference is that the grouping is for “abstract entities or concepts.” This ST seems to be misused in the UMLS, probably due to confusion about its nature. The UMLS User Note (UN) for this ST says “Few concepts will be assigned to this broad type.” Nevertheless as many as 609 concepts are assigned this ST. For comparison, there are only 19, 45, and 115 concepts assigned the broad STs, **Entity**, **Physical Object**, and **Event**. In unpublished research of our group, 11 concepts, out of a randomly selected sample of 50 concepts with **CE** assignments, were judged to need a more specific ST, namely a descendant of **CE**, rather than **CE**. Six more concepts of the sample should have been assigned other STs, which are not descendants of **CE**.

The confusion concerning the two STs **Classification** and **CE** is easily illustrated in the META subhierarchy rooted at *Anatomical term,* assigned **CE**. Its parent *Non-physical anatomical entity* is assigned **Classification** and so is its child *General anatomical entity*. The other three children, *Embryological term, Histological term*, and *Radiological term*, are

assigned **CE**. The last of these children is a leaf (has no children) in spite of the purpose of **CE** to model groupings of concepts. Three of the siblings of *Anatomical term* are also assigned **CE**. The child *General anatomical entity* has many children and descendants. About 20 of the children are assigned **Classification** and most of those have no children at all, in spite of the intention that **Classification** should model a group. There does not seem to be a clear distinction between the concepts in this META subhierarchy assigned **CE** and those assigned **Classification**. These assignments seem to be used interchangeably. Hence, a concept that is assigned **CE** may also serve as an obstacle for the expansion methodology of [8].

6 Conclusion

We have presented the solution to a major problem encountered by a previous algorithm [8] for finding concepts likely to be lacking a semantic type assignment. It was recognized that the problem was caused by (chains of) concepts assigned the “unusual” semantic type **Classification,** which interrupted the expansion process of the previous algorithm [8]. On the other hand, “uncontrolled” expansion of the algorithm would have led to an undesirable explosion of the number of concepts requiring human auditing.

Thus we presented and justified the design of a new expansion methodology with obstacle avoidance (EM/OA) and showed that this methodology successfully discovered over 500 concepts lacking the assignment of the semantic type **Experimental Model of Disease.** We also demonstrated other semantic types besides **EMD** for which the EM/OA can be successfully applied. As the lack of semantic types is often indicative of other errors [7, 13, 14], the importance of this algorithm goes beyond identifying missing ST assignments.

Acknowledgments

This work was partially supported by the NLM under grant R-01-LM008445-01A2 and by the Department of Health and Human Services grants 3R01LM008445-03S1 and 3R01LM008445-03S2.

REFERENCES

[1]. Bodenreider, O. Circular hierarchical relationships in the UMLS: Etiology, diagnosis, treatment, complications and prevention. Proc. 2001 AMIA Annual Symposium; 2001. p. 57-61.

[2]. Bodenreider, O. Strength in numbers: Exploring redundancy in hierarchical relations across biomedical terminologies. Proc. 2003 AMIA Annual Symposium; 2003. p. 101-105.

[3]. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research. 2004; 32(D):267–270.

[4]. Campbell KE, Oliver DE, Shortliffe EH. The Unified Medical Language System: Toward a collaborative approach for solving terminologic problems. Journal of the American Medical Informatics Association. 1998; 5(1):12–16. [PubMed: 9452982]

[5]. Chen Y, Perl Y, Geller J, Cimino JJ. Analysis of a study of the users, uses and future agenda of the UMLS. Journal of the American Medical Informatics Association. 2007; 14(2):221–231. [PubMed: 17213497]

[6]. Chen Y, Gu H, Perl Y, Geller J, Halper M. Structural group auditing of a UMLS semantic type’s extent. Journal of Biomedical Informatics. February; 2009 42(1):41–52. [PubMed: 18619563]

[7]. Chen Y, Gu H, Perl Y, Geller J. Structural group-based auditing of missing hierarchical relationships in UMLS. Journal of Biomedical Informatics. June; 2009 42(3):452–467. [PubMed: 18824248]

[8]. Chen Y, Gu H, Perl Y, Halper M, Xu J. Expanding the extent of a UMLS semantic type via group neighborhood auditing. Journal of the American Medical Informatics Association. September/ October; 2009 16(5):746–757. [PubMed: 19567802]

[9]. Cimino JJ. Auditing the Unified Medical Language System with semantic methods. Journal of the American Medical Informatics Association. 1998; 5:41–51. [PubMed: 9452984]

[10]. Cimino, JJ. Battling Scylla and Charybdis: the search for redundancy and ambiguity in the 2001 UMLS Metathesaurus. In: Bakken, S., editor. Proc. 2001 AMIA Annual Symposium; 2001. p. 120-124.

[11]. Geller J, Gu H, Perl Y, Halper M. Semantic refinement and error correction in large terminological knowledge bases. Data and Knowledge Engineering. 2003; 45(1):1–32.

[12]. Gu H, Perl Y, Geller J, Halper M, Liu LM, Cimino JJ. Representing the UMLS as an OODB: Modeling issues and advantages. Journal of the American Medical Informatics Association. Jan-Feb; 2000 7(1):66–80. [PubMed: 10641964] Selected for reprint in: Haux R, Kulikowski C. Yearbook of Medical Informatics: Digital Libraries and Medicine. 2001:271–285.SchattauerStuttgart, Germany(International Medical Informatics Association)

[13]. Gu H, Perl Y, Elhanan G, Min H, Zhang L, Peng Y. Auditing concept categorizations in the UMLS. Artificial Intelligence in Medicine. May; 2004 31(1):29–44. [PubMed: 15182845]

[14]. Gu, H.; Hripcsak, G.; Chen, Y.; Morrey, CP.; Elhanan, G.; Cimino, JJ.; Geller, J.; Perl, Y. Evaluation of a UMLS auditing process of semantic type assignments. In: Teich, JM.; Suermondt, J.; Hripcsak, G., editors. Proc. 2007 AMIA Annual Symposium; 2007. p. 294-298.

[15]. Humphreys BL, Lindberg DAB, Schoolman HM, Barnett GO. The Unified Medical Language System: An informatics research collaboration. Journal of the American Medical Informatics Association. 1998; 5(1):1–11. [PubMed: 9452981]

[16]. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Methods of Information in Medicine. 1993; 32:281–291. [PubMed: 8412823]

[17]. McCray AT, Hole WT. The scope and structure of the first version of the UMLS Semantic Network. Proc. Fourteenth Annual SCAMC. 1990:126–130.

[18]. McCray AT. An Upper-Level Ontology for the Biomedical Domain. Comparative and Functional Genomics. 2003; 4:80–84. [PubMed: 18629109]

[19]. Morrey CP, Geller J, Halper M, Perl Y. The Neighborhood Auditing Tool: a hybrid interface for auditing the UMLS. Journal of Biomedical Informatics. June; 2009 42(3):468–489. [PubMed: 19475725]

[20]. Mougin, F.; Bodenreider, O. Approaches to eliminating cycles in the UMLS Metathesaurus: Naïve vs. formal. Proc. 2005 AMIA Annual Symposium; 2005. p. 550-554.

[21]. Mougin, F.; Bodenreider, O. Auditing the NCI Thesaurus with Semantic Web Technologies. Proc. 2008 AMIA Annual Symposium; 2008. p. 500-504.

[22]. [accessed in October 2007] NCIt Thesaurus. http://nciterms.nci.nih.gov/NCIBrowser/Dictionary.do

[23]. Peng, Y.; Halper, M.; Perl, Y.; Geller, J. Auditing the UMLS for redundant classifications. Proc. 2002 AMIA Annual Symposium; 2002. p. 612-616.

[24]. Schuyler PL, Hole WT, Tuttle MS, Sherertz DD. The UMLS Metathesaurus: Representing different views of biomedical concepts. Bulletin of the Medical Library Association. 1993; 81(2): 217–222. [PubMed: 8472007]

[25]. Tuttle, MS.; Sherertz, DD.; Olson, NE.; Erlbaum, MS.; Sperzel, WD.; Fuller, LF., et al. Using META-1, the first version of the UMLS Metathesaurus. Proc. Fourteenth Annual SCAMC; 1990. p. 131-135.

[26]. Zhu X, Fan JW, Baorto D, Weng C, Cimino JJ. A Review of Auditing Methods Applied to the Content of Controlled Biomedical Terminologies. Journal of Biomedical Informatics. June; 2009 42(3):413–425. [PubMed: 19285571]

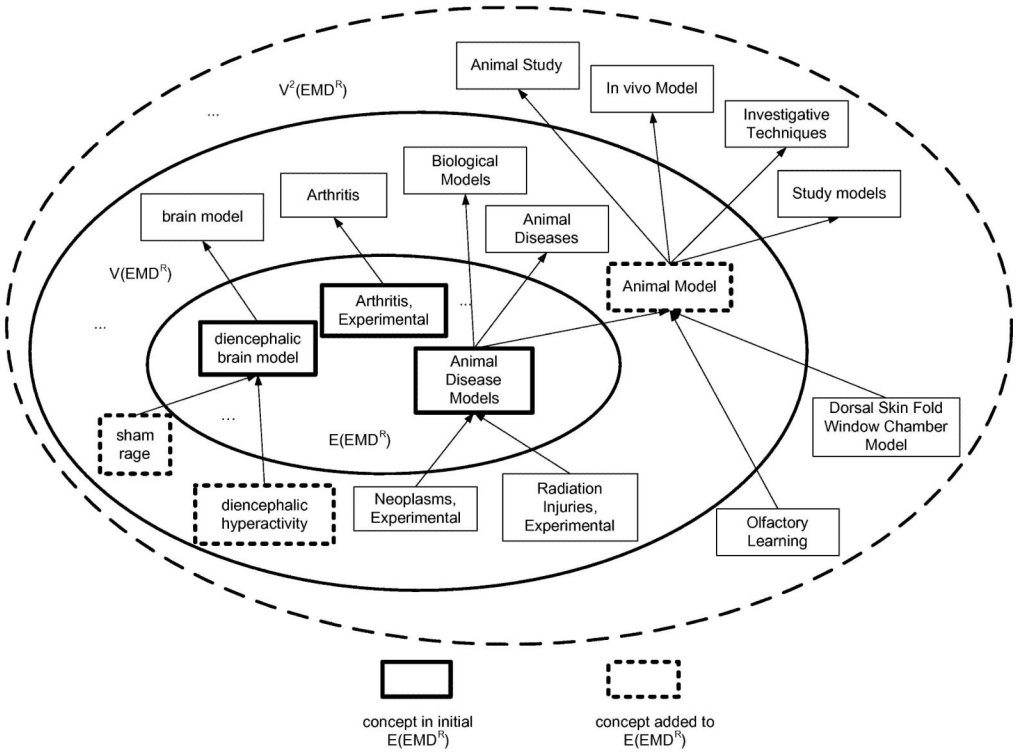

Figure 1.
Auditing the RST **EMD**^{R}

Figure 2.
Processing of AUX(**EMD**^{R})

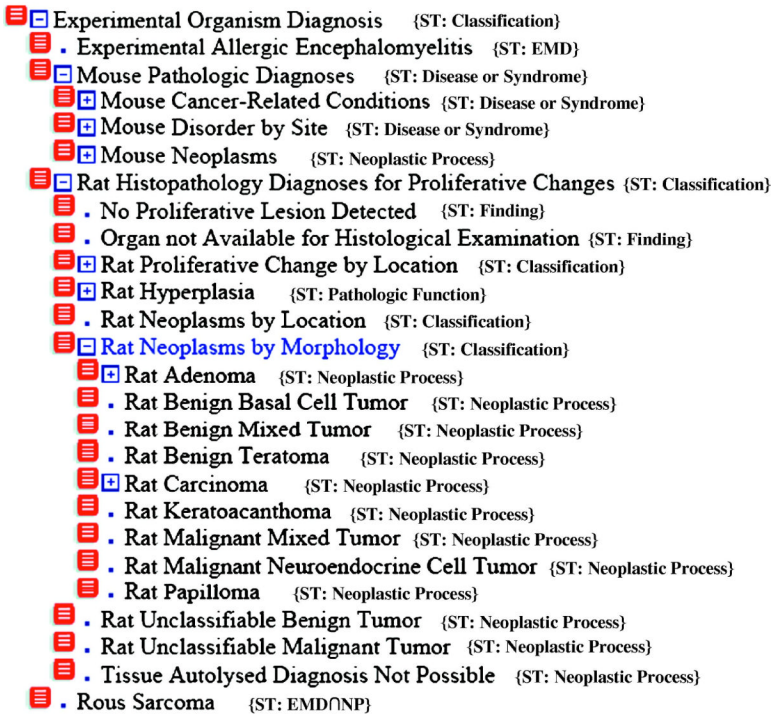

Figure 3.
Partial Indented Experimental Organism Diagnosis Hierarchy

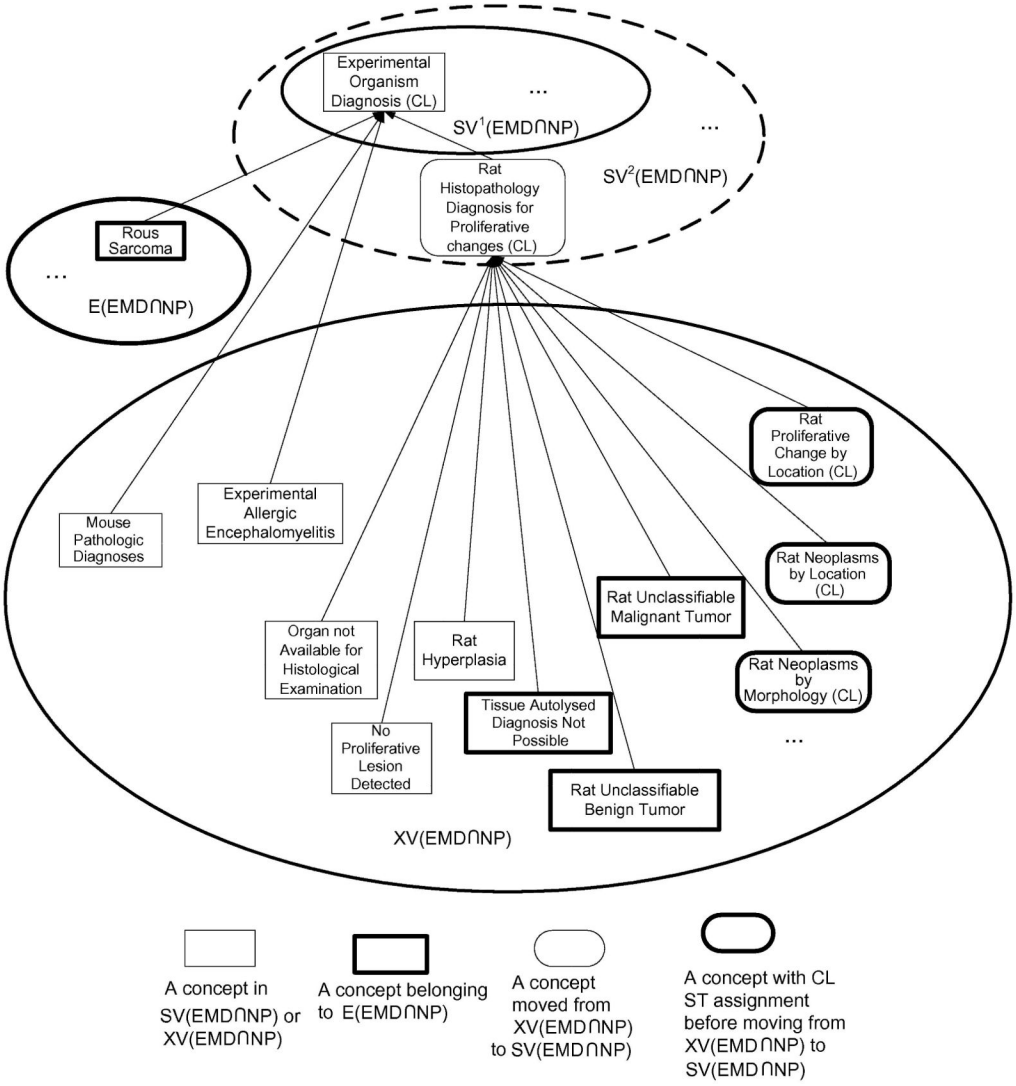

Figure 4.

Applying the Expansion Methodology with Obstacle Avoidance to Pass Over Obstacles (p=3 and i=2)

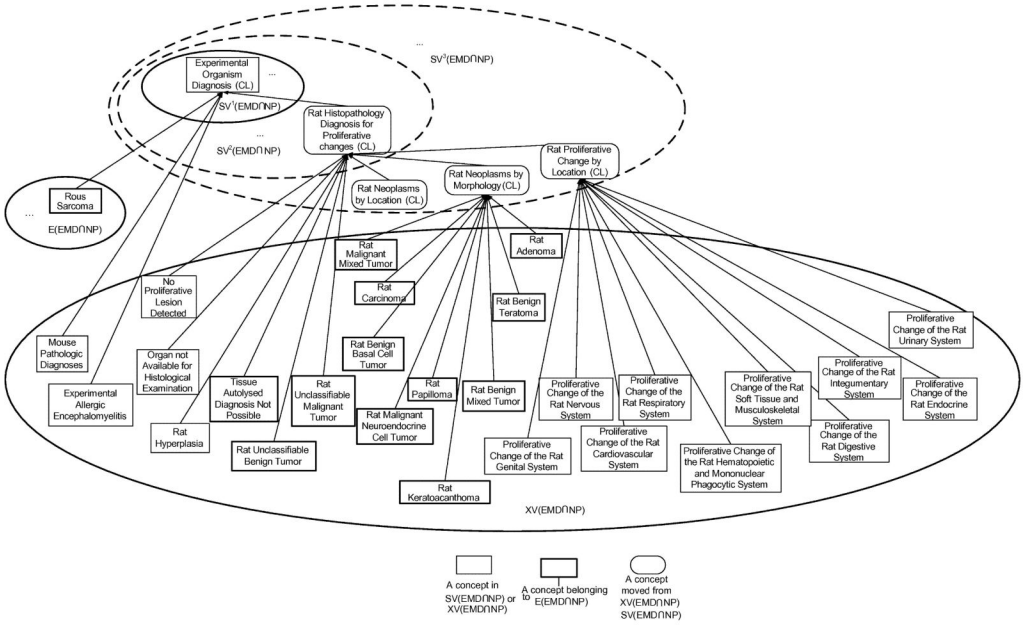

Figure 5.
Applying the Expansion Methodology with Obstacle Avoidance to Byass Obstacles (p=3 and i=3)

Table 1

Results of processing of envelopes of **EMD∩NP** (New assignments starting at Stages 1-3)

Envelope	# Cpt in envelope	# Added to E(EMD ∩ NP)	Error Rate (%)	E(EMD ∩ NP)
V	33	9	27	42
V^2	44	26	59	68
V^3	79	78	99	146
V^4	212	201	95	347
V^5	214	204	95	551
V^6	137	135	99	686
V^7	145	119	83	805
V^8	97	92	95	897
V^9	32	32	97	929
V^{10}	17	17	100	946
V^{11}	2	2	100	948
V^{12}	–	–	–	948
Stages 1-3	20	12	60	960
V^{13}	20	19	95	979
V^{14}	5	5	100	984
V^{15}	20	19	95	1003
V^{16}	14	13	93	1016
V^{17}	62	62	100	1078
V^{18}	5	5	100	1083
V^{19}	–	–	–	1083
Total:	146	135	92	1083

Table 2

Results of processing envelopes of **EMD**

Stage	#Concepts in SV (EMD)	#Concepts in XV (EMD)	#Concepts in envelope	#Concepts Added to E(EMD)	Error Rate (%)	E(EMD)
Stage 1	1	4	–	–	–	165
Stage 2	1	10	–	1	10	166
Stage 3	3	21	43	52	81	218
V^1	–	–	74	62	84	280
V^2	–	–	–	–	–	280
Total:	–	**35**	**117**	**115**	**76**	**280**

Table 3

Results of processing envelopes of **EMD∩NP** by applying Part 2 of the methodology [8]

Envelope	#Concepts in envelope	#Concepts Added to E(EMD∩NP)	Error Rate (%)	E(EMD∩NP)|
V^1	232	222	96	1315
V^2	82	82	100	1397
V^3	–	–	–	1397
Total:	**314**	**304**	**97**	**1397**