Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Biomedical Informatics 37 (2004) 77–85
Journal of Biomedical Informatics
www.elsevier.com/locate/yjbin

Towards the development of a conceptual distance metric for the UMLS

Jorge E. Caviedes[a,*] and James J. Cimino[b]

[a] Intel Corporation, 5000 W. Chandler Blvd., Chandler, AZ 85226, USA
[b] Department of Biomedical Informatics, Columbia University, New York, NY, USA

Received 20 October 2003

Abstract

The objective of this work is to investigate the feasibility of conceptual similarity metrics in the framework of the Unified Medical Language System (UMLS). We have investigated an approach based on the minimum number of parent links between concepts, and evaluated its performance relative to human expert estimates on three sets of concepts for three terminologies within the UMLS (i.e., MeSH, ICD9CM, and SNOMED). The resulting quantitative metric enables computer-based applications that use decision thresholds and approximate matching criteria. The proposed conceptual matching supports problem solving and inferencing (using high-level, generic concepts) based on readily available data (typically represented as low-level, specific concepts). Through the identification of semantically similar concepts, conceptual matching also enables reasoning in the absence of exact, or even approximate, lexical matching. Finally, conceptual matching is relevant for terminology development and maintenance, machine learning research, decision support system development, and data mining research in biomedical informatics and other fields.
© 2004 Elsevier Inc. All rights reserved.

Keywords: Conceptual metrics; Similarity metrics; Semantic distance; UMLS; Medical terminology

1. Introduction

The Unified Medical Language System (UMLS) is a knowledge representation framework designed to support broad scope biomedical research queries. It includes over 100 medical terminology sources (i.e., narrow-scope sets of concepts designed to be used in specific medical domains), as well as a variety of imported and native semantic and syntactic structures [1]. The two major resources of the UMLS are the Metathesaurus, which contains a large collection of concepts, and the Semantic Network (SN), which contains semantic types that form an abstraction of the Metathesaurus (the Metathesaurus and the SN are distributed in formatted text files such as MRSO, which includes terminology sources of concepts, and MRREL, which includes relationships among concepts). Although lexical matching of terms to UMLS concepts, mapping of free text to UMLS concepts [2], and some UMLS-based applications that allow natural language input (e.g., SAPHIRE [3]) are presently supported, semantic concept matching is not generally available.

Conceptual matching is an essential component of human and machine reasoning as it enables applying problem-solving principles and making inferences (using high-level, generic concepts), based on available data (typically represented as low-level, specific concepts). A conceptual matching metric gives a quantitative similarity score between two concepts, by definition a value of zero corresponds to identical concepts, while a large score means the concepts are very different. Through the identification of semantically similar concepts, conceptual matching enables reasoning in the absence of exact, or even approximate, lexical matching. Because of this ability to match abstract concepts, conceptual matching is relevant for terminology development and maintenance, machine learning, decision support systems, and data mining research in biomedical informatics and other fields. Prior work on conceptual matching can be found in the areas of knowledge-based representations,

* Corresponding author. Fax 1-480-554-4880.
E-mail address: jorge.e.caviedes@intel.com (J.E. Caviedes).

1532-0464/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2004.02.001

hypermedia document search [4,5], and more recently in multimedia database research.

To be computationally useful, conceptual matching must be expressed as a quantitative metric that can be used to set decision thresholds or apply approximate matching principles. Given that concepts are not necessarily atomic units, but may be expressed as structured sets of other concepts, work on conceptual metrics must address distance between concepts as well as distance between sets of concepts.

In the UMLS, the concept representations are not homogeneous, sometimes inconsistent, and may be incompatible. This poses a problem for its use in the development of conceptual distance metrics. We are aware that the UMLS contents are largely hand coded and maybe inadequate in some instances. However, the strong interest and steady progress towards meeting formal terminology requirements, first set forward in the late 1990s by Campbell and others (see [6,7]), provide a reasonable basis for using the UMLS in the present work. Moreover, through this work we also hope to emphasize the advantages of improving coverage and consistency of concept-oriented representations.

If we assume that there is sufficient semantic content, i.e., coverage, and that it can be accessed in the UMLS, these are the main issues:

- How to assess similarity between concepts within and across chosen UMLS terminology sources, and
- How to assess similarity between concepts independent of terminology.

The ability to compute conceptual distance between concepts within a given terminology in turn enables the following capabilities for retrieval applications:

- Post-processing of lexical matching results by filtering out irrelevant concepts found before the target concept is searched in the actual documents or records.
- Ranking matching documents or records by conceptual distance to the target concept(s) in a bibliographic search.

Also within a given terminology, conceptual metrics enable the following capabilities for terminology development and maintenance:

- Detection of redundancies (or coding errors) if non-synonym terms for different concepts are found to have zero distance between them, and
- Detection of false similarities or dissimilarities that may suggest errors in the conceptual structures.

And, when conceptual metrics are applied across terminologies they enable:

- Investigating the conceptual space (or scale) and its consistency across terminologies, i.e., similarity rankings should be preserved across terminologies.
- Making inferences about seamless merging of terminologies based on whether monotonicity of the conceptual similarity rankings is preserved in multi-hierarchical structures.

In this paper, we review general conceptual and lexical matching principles, and present an algorithm for conceptual metrics in the UMLS. We have applied our metric to some sampled domains and present the results of our evaluation.

2. Semantic and lexical matching principles

Concepts in the UMLS belong to a semantic network and hierarchical structures [8]. A natural consequence of that design is that similar concepts should be close to each other. The first published work on a metric for MeSH terms in the UMLS framework was carried out by Rada [9]; however, it used all *broader-than* (called "related-broader in the UMLS, or RB for short) links in lieu of *is-a* links, it did not deal with the case of single or multiple terminology sources and did not address the imperfect conceptual structures of the UMLS. To the best of our knowledge, after the initial work by Rada, similar or related work using the current version of the UMLS has not been published.

The results obtained by Rada, using an earlier version of the UMLS, indicated a high potential for a metric based on the minimum path along RB, links. However, in the present version of the UMLS, RB links are too numerous, do not span a consistent topological concept graph, and do not enforce consistency in any particular terminology source. The main advantage of a concept-to-concept metric such as the one proposed by Rada is that it is a true metric; i.e., $d(C_i, C_j)$ the distance between concepts C_i and C_j has the following properties:

1. Non-negative definiteness

$$d(C_1, C_2) \geqslant 0, \text{ and } \quad d(C_1, C_2) = 0 \text{ iff } C_1 = C_2 \tag{1}$$

2. Symmetry

$$d(C_1, C_2) = d(C_2, C_1) \tag{2}$$

3. Triangular inequality

$$d(C_1, C_2) \leqslant d(C_1, C_3) + d(C_3, C_2) \tag{3}$$

Although, symmetry is a controversial issue, as some researchers argue it should hold [10] and some argue against it [11], these three properties are very important for logical and computational tractability. Other Euclidean metrics based on geometric distances in a feature space (the set of quantifiable properties of a concept), as proposed in the statistical pattern recognition literature, and currently applied to content based retrieval of video and multimedia, e.g. [12], are possible but very likely too computationally expensive for practical use.

Published work on lexical matching encompasses the methods currently used for document retrieval, either by indexed keywords, or by full document search does not generally address conceptual matching. In those works, documents are indexed by exact or inexact keyword matching, and retrieval is based on logical combinations

of the matching sets. Aalbersberg has applied information theory principles such as the law of Zipf (i.e., term frequency is inversely proportional to information content so that the product of frequency times sorting rank is approximately constant) to the document retrieval problem [13]. Another method, which has been applied to case based reasoning systems, but does not use term frequency, allows inexact matching and uses a measure of entropy between strings (i.e., increasing disorder in individual words and letters increases distance between free text strings) has been proposed by Caviedes [14].

As the semantic principles of concept orientation, thesauri, and conceptual structures become more popular, the gap between lexical and conceptual matching has started to close, mainly driven by the need to support intelligent matching that can overcome the obvious brittleness of lexical matching. A notable example in this category is WordNet, a lexical database developed by the Cognitive Science Laboratory at Princeton University, in which groups of synonyms, connected by various types of links between themselves, are used to represent the same underlying lexical concept [15]. Also of interest, in the area of multimedia content retrieval, Kazman et al. [12] have used lexical trees built from the terms extracted from the audio stream using speech recognition.

3. Requirements and algorithm for conceptual metrics in the umls

By design, the UMLS offers flexibility to work with any terminology of choice. Therefore, conceptual metrics of practical interest require:

- Metrics for concepts within a single terminology, which are useful for applications limited to one source within and outside the UMLS.
- Metrics for multiple chosen terminologies, where the conceptual distances for the same concepts in different and joint sources may shed light on the appropriateness of the mapping across terminologies.
- Generic metrics for UMLS concepts, to compute conceptual distance between any two concepts without specifying terminology.

Given that the hierarchies spanned by UMLS links are not guaranteed to stay within a chosen terminology (or terminologies), we deal with the first two requirements using hierarchies spanned by the *parent* (PAR) links within the selected terminologies. PAR links are semantically similar to is-a links and are subsumed by RB links (i.e., parent links are normally included in broader-than links). In order to facilitate the search for paths between concepts, it is possible to assume that the PAR hierarchies are *directed acyclic graphs* (DAGs) and discard cycles as inconsistencies. For the third requirement, we investigate the use of PAR and RB links.

The presence of cycles in the PAR or RB hierarchies merits further analysis. Simply discarding cycles (caused presumably but one or more incorrect links) may lead to erroneous distance values. This would happen if the shortest path includes the incorrect link. Therefore, maintaining the search within terminologies known to include verifiable DAG hierarchies should be the preferred technique for future applications based on the proposed method. Additionally, the method could report the distances in which cycles are involved as cases to be debugged. Further research may lead to robust methods to deal with graph cycles especially if the underlying problem can be traced to provable syntactic/semantic inconsistencies.

To address the requirements stated above, we have devised and implemented an algorithm to compute the conceptual distance between pairs of concepts in the UMLS using PAR links within one or more terminologies, and using RB links for the more general case.

An example of concepts connected by PAR links is shown in Fig. 1. The shortest path along PAR links is clearly associated with the similarity between pairs of concepts. The terminology source (SRC) field in the UMLS Metathesaurus MRREL file can be used. The MRREL information for RB links has been found to be less reliable for RB links than PAR links. Thus, the RB links must be extracted from the MRSO file, which is a slower process.

The main steps of our algorithm include first verifying the terminology source of the input CUIs (numerical codes or Concept Unique Identifiers cui1 and cui2). Then, get the PAR or RB sub-graphs (these are called the spread activation graphs or SAGs). And, finally finding shared CUIs and the minimum path avoiding circular and infinite paths; the minimum path is the conceptual distance or *CDist*. The flowchart is shown in Fig. 2.

4. Methods

The objective of the experimental method is to assess the feasibility of the conceptual distance approach by comparing it against subjective similarity estimates by domain experts. The two main cases to be investigated are distance between individual concepts, and distance between sets of concepts (i.e., concept clusters). The method is summarized in Table 1.

Some applications of conceptual distance will ultimately be used to rank concepts. Based on the mathematical principles presented in Section 2, it is appropriate to give the conceptual distance function and space a fully parametric treatment (including its statistical validation). We have therefore been careful to elicit numerical (not categorical) subjective distance estimates, not ranks. This approach simplifies data collection as well as

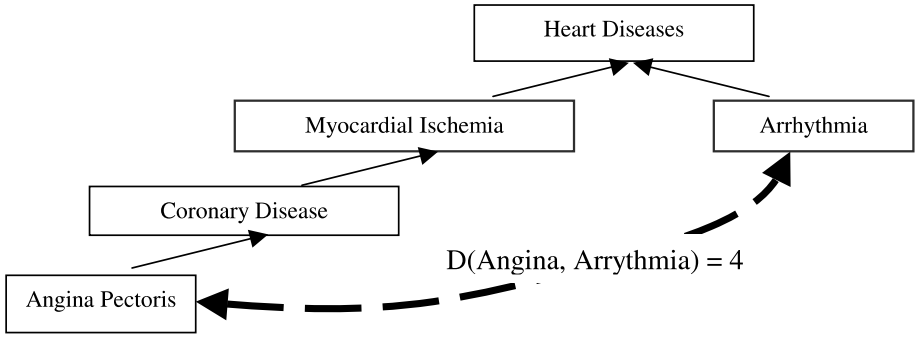

Fig. 1. The shortest path along PAR/RB links is closely related to their conceptual distance.

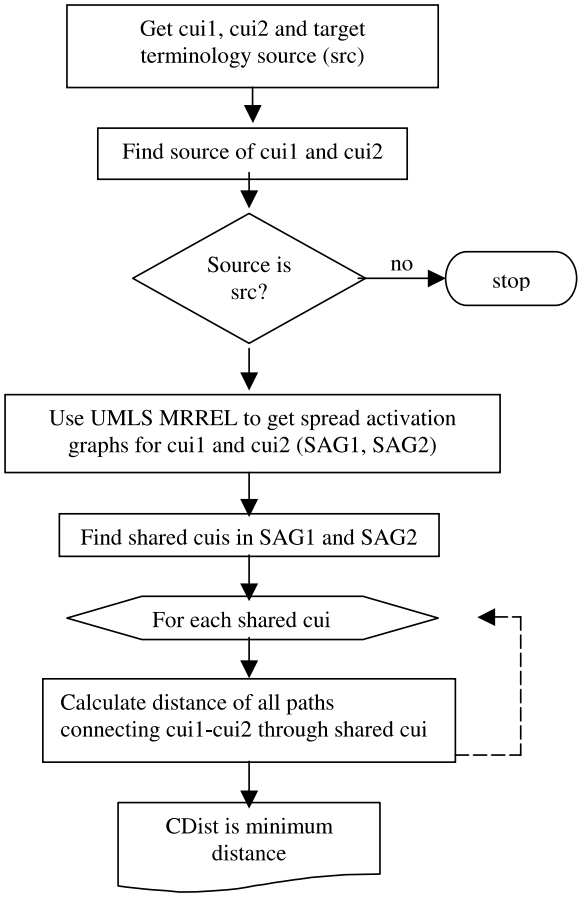

Fig. 2. Flowchart of the CDist algorithm. *Cui1* and *cui1* are concept unique identifiers, and src is the terminology source.

statistical interpretation. Non-parametric statistics such as Spearman's correlation coefficient and Kendall's coefficient of concordance could be applied if we converted our parametric data to ranked data and checked performance against conceptual distance and degree of agreement among experts respectively (Cohen's Kappa would not apply as we are not evaluating similarity in a dichotomous output case). However, these techniques would not be straightforward to apply in our case due to their limitation when dealing with tied ranks (tied conceptual distance values are found quite often).

Our initial statistical model is simple, it is only aimed at finding evidence to suggest significant correlation between reference expert scores and the conceptual distance, provided that there is enough agreement among the experts (e.g., low standard deviation for the set of subjective estimates of each score). A rigorous factorial ANOVA for all possible distances is not possible at the present stage given that we only have subjective scores from three experts. However, it would be advisable when more data becomes available.

Validation against average subjective estimates using several concept sets and terminologies appears to be an appropriate way to test relevance of conceptual distance to literature search and knowledge-based applications. More formal tests are not possible at this time because there are no standard concept sets and associated distances that could be used for benchmarking.

Three sets of concepts, including a set of concept clusters, were used to study concept-to-concept similarity and cluster-to-cluster similarity. The concepts were chosen without restricting their semantic or taxonomic type, and include symptoms, diseases, pathologies, and anatomic and physiological concepts. The concept clusters are groups of three related concepts each, i.e., they could be associated with a clinical situation, and illustrate the need to match those situations to others in a database or the literature. At this time we simply consider the co-occurrence of the concepts in the clusters, without concern for how they are related to each other. This leads to the simplest cluster similarity approach.

The domain experts are physicians, who were asked to score subjective similarity between concepts or clusters using a numerical scale from zero to any arbitrary maximum of their choice. Involving three physicians in the initial test provided enough data to analyze agreement with conceptual distance, and issues such as overall variance and specific cases of large disagreement. Given the large extent of agreement observed so far, we believe that involving more than 5 experts in future tests may not be necessary.

The following tests have been carried out:

Table 1
Summary of method to validate conceptual distance

Objective	Description	Validation
Investigate distance between a pair of concepts	Compute conceptual distance between pairs of concepts for similar, less similar, and unrelated concepts	Compare against average of subjective estimates of conceptual distance by domain experts
Investigate distance between concept clusters	Calculate inter-cluster distance using one of several methods, e.g., average of all possible distances across two clusters	Compare against average of subjective estimates of cluster distances by domain experts

1. Conceptual distances within three terminologies (MSH, SNMI, and ICD9CM)[1] for a set of 11 concepts. We computed all CDist values and compared them against the average subjective scores provided by three physicians. For convenience, expert scores provided by physicians and CDist values, have been normalized by the maximum value found in each test (norm_score = original_score * max_score_of_test/ max_score_used_by_this_expert).
2. Conceptual distances among unrelated concepts. We analyzed CDist and expert scores for a set of 11 mostly unrelated concepts. The data are expert scores for all distances to one concept, and all CDist values.
3. Conceptual distances among clusters of concepts. For a set of 4 clusters of 3 concepts each, we study CDist within clusters and across clusters using different terminologies.
4. Using RB links to compute conceptual distance. We have tested the feasibility of using unrestricted RB links to compute CDist instead of PAR links.

5. Results

5.1. Conceptual distance within three terminologies

For a set of 11 concepts shown in Table 2, we have computed 55 possible CDist values (or less if the concepts were not found in the terminology source) within MSH, SNMI, ICD9CM, and joint MSH-SNMI. Next we obtained the expert scores and compared their average values against the CDist values.

The scatter plots for the three CDist cases (including standard deviation bars for CDist-MSH) are shown in Fig. 3, and the correlations between the CDist values and average expert scores are presented in Table 3.

CDist in MSH shows the highest correlation with the expert scores. SNMI had the lowest correlation with the expert scores (0.6), and ICD9CM was the second best (0.74). MSH joined with SNMI still did better than all others except MSH (0.75), while using RB links restricted to MSH terms did second to last (0.68); considering this drop in performance and the fact that two distances that were found for MSH could not be found for MSH-SNMI, this suggests that RB links may not include all PAR links. Suppressing the 3 expert scores with the largest standard deviations (associated with lack of consensus among experts) leads to increases of up to 0.04 in the correlations (see numbers in parenthesis in Table 3).

Table 2
Set of concepts tested in within specific terminologies

CUI	Name
C0002962	Angina pectoris
C0003811	Arrhythmia
C0007192	Cardiomyopathy, alcoholic
C0010068	Coronary disease
C0018799	Heart diseases
C0018834	Heartburn
C0018802	Heart failure, congestive
C0000737	Abdominal pain, unspecified site
C0020621	Hypokalemia
C0030631	Passive-aggressive personality disorder
C0035238	Respiratory system abnormalities

Table 4 shows the correlations between CDist in MSH and each of the other CDist values. ICD9CM is the single most consistent terminology with MSH, while those that include MSH are better than SNMI or ICD9CM by themselves. Also, notice that although the RB/MSH CDist (i.e., use RB links constrained to MSH concepts) is the second closest to MSH, as indicated before, it did not show a high correlation with the expert scores.

The largest prediction errors, between CDist and the expert scores, were observed for the distances:

- Angina to Abdominal Pain,
- Respiratory Abnormalities to Arrhythmia, and
- Heart Diseases to Passive Aggressive Personality Disorder

In these three cases, CDist largely underestimated the similarity. The paths, CDist, expert scores, and *depth*, i.e., shortest path from most specific common ancestor to a root concept are the following:

[1] MSH is the Medical Subject Headings standard terminology. SNMI is Systematized Nomenclature of Medicine, or SNOMED. ICD9CM is the International Classification of Diseases, ninth revision, Clinical Modifications.

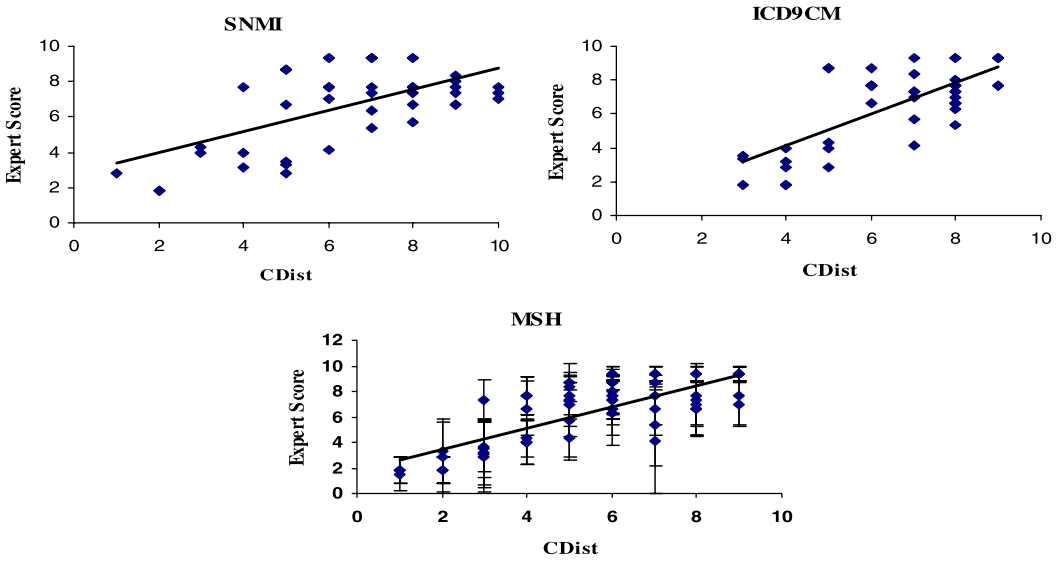

Fig. 3. CDist vs. average expert scores for three terminologies.

C0002962 **Angina Pectoris**
C0008031 Chest Pain
–C0030193 ***Pain***
C0000737 **Abdominal Pain**

The depth of concept ***Pain*** is 6, CDist between **Angina Pectoris** and **Abdominal Pain** is 3, AVG-Score = 7.3

C0035238 **Respiratory System Abnormalities**
C0035242 Respiratory Tract Diseases
–C0012674 ***Diseases***
C0039058 Pathological Conditions, Signs and Symptoms
C0030660 Pathologic Processes
C0003811 **Arrhythmia**

The depth of ***Diseases*** is 3, CDist between **Respiratory System Abnormalities** and **Arrhythmia** is 5, AVG-Score = 8.6

C0018799 **Heart Diseases**
C0007222 Cardiovascular Diseases
C0012674 Diseases (MeSH Category)
–C1135589 ***Index Medicus Descriptor***
C0033874 Psychiatry and Psychology (MeSH Category)
C0004936 Mental Disorders
C0031212 Personality Disorders
C0030631 **Passive-Aggressive Personality Disorder**

Depth of ***Index Medicus*** is 2, CDist **Heart Diseases** to **Passive-Aggressive Disorder** is 7, AVG-Score = 9.33

The large differences between CDist's and the expert scores above could be explained as comparisons involving very general concepts. We could argue that CDist values being equal, pairs of closely related general concepts are less similar than pairs of more specific concepts. The last two cases above would suggest that making the conceptual distance inversely proportional to the depth would improve performance (although in the first case this would not help, it would not degrade it either because it was found to be at the average depth). In related work by Tudhope and Taylor, which also uses a shortest path method [5], they use the level in the hierarchy to define a *specialization factor*, which can be controlled by a chosen weight so that siblings at the top of the hierarchy result less similar than siblings at the bottom.

Table 3
Correlations between CDist and expert scores

Scores	Correlation
CDist (MSH) vs. Expert-Score	0.77 (0.81)
CDist(MSH&SNMI) vs. Expert-Score	0.75 (0.79)
CDist (ICD9CM) vs. Expert-Score	0.74 (0.78)
CDist (MSH/RB) vs. Expert-Score	0.68 (0.75)
CDist (SNMI) vs. Expert-Score	0.60 (0.62)

Table 4
Consistency of other CDist values with those within MSH

Scores	Correlation
CDist(MSH) vs. CDist(MSH&SNMI)	0.90
CDist(MSH) vs. CDist(RB/MSH)	0.83
CDist(MSH) vs. CDist(ICD9CM)	0.78
CDist(MSH) vs. CDist(SNMI)	0.50

5.2. Conceptual distances among unrelated concepts

Given that conceptual distance can be calculated also for dissimilar concepts, we conducted a second test as an early attempt to study the behavior of the metric outside the normal range, and to see if there is any meaning associated to the distance that could be assigned to unrelated concepts. For the set of 11 mostly unrelated

Table 5
Set of mostly unrelated concepts tested

CUI	Name
C0878705	Plica syndrome
C0878707	Precipitous drop in hematocrit
C0878752	Abnormal loss of weight and underweight
C0878754	Genetic counseling and testing on procreative management
C0878756	Perpetrator of child and adult abuse
C0878757	Child battering and other maltreatment by father, stepfather, or boyfriend
C0917805	Transient cerebral ischemia
C0917967	Pupillary functions, abnormal
C0920296	Reading disorder, developmental
C0936250	Eczema herpeticum
C0949122	Acute laryngitis without mention of obstruction

concepts shown in Table 5, all the CDist values were obtained (for ICD9CM), and also two sets of expert scores from the first concept to all the others.

We obtained values of 7 or larger for CDist, and 6.7 or larger for the expert scores. We noticed that the closest concepts, C0878756 and C0878757 (child abuse concepts) lay at equal distance from the first concept, C0878705, and the distance between them is the smallest (CDist is 1).

We also carried out a test for a set with five concepts (including a mix of anatomical, disease, and treatment concepts) shown in Table 6. All possible CDist values for SNMI and MSH sources are shown in Table 7. The predominantly large CDist values confirm strong dissimilarity in all cases, except for the distance between the first two concepts, as one may expect. (No expert scores were used in this case.)

The correlation between the SNMI and MSH scores above is 0.78. This suggests that conceptual similarity, including unrelated concepts, could be used to check consistency between terminologies, e.g., based on conceptual distance agreement, and to verify for incomplete or incorrect hierarchies.

5.3. Conceptual distance among clusters of concepts

Four clusters of three concepts each have been used to investigate conceptual distance between concepts that can be expressed as sets of other concepts, or clusters (notice that this does not imply tight clusters, i.e., distances within clusters do not have to be smaller than distances across clusters). Table 8 shows the three possible concept pairs within each cluster.

Table 6
Set of SNMI concepts tested

CUI	Name
C0012242	Digestive system diseases
C0014869	Peptic esophagitis
C0033968	Psychotherapy
C0039971	Thirst
C0039979	Thoracic duct

Table 7
CDist values for all distances in the set of SNMI concepts

CUIs	SNMI CDist	MSH CDist
C0012242 C0014869	3	3
C0012242 C0039971	6	7
C0012242 C0033968	7	5
C0012242 C0039979	7	7
C0014869 C0039971	9	10
C0033968 C0039971	9	6
C0039971 C0039979	9	10
C0014869 C0033968	10	8
C0014869 C0039979	10	10
C0033968 C0039979	10	8

We have computed CDist using MSH and SNMI for all pairs of concepts within each cluster (3 distances within each cluster) and all possible pairs of concepts across clusters (9 distances for each of the 6 possible cluster pairs). We also obtained expert scores from three physicians for all pairs within each cluster and the 6 possible distances between clusters.

The average distances within the clusters are summarized in Table 9. The results excluding the two largest prediction errors (distances Thirst-Mouth, and Enteritis-Intestines) are shown in parenthesis. These errors were interesting because, for the experts those were the smallest distances, while for CDist they were as dissimilar as all the others.

The results for cluster-to-cluster distances are summarized in Table 10. Although the range and number of distances is not very large, and the test does not consider concept structures in the clusters (which are likely to play a role in the matching operation), the fit between CDist and expert score is very encouraging as evidenced

Table 8
Clusters of concepts used to test intra-cluster distance

CUI name pair	CUI pair
Cluster 1	
Bacteria–Gastritis	C0004611–C0017152
Gastritis–Esophagus	C0017152–C0014876
Esophagus–Bacteria	C0014876–C0004611
Cluster 2	
Virus–Enteritis	C0042776–C0014335
Enteritis–Intestines	C0014335–C0021853
Intestines–Virus	C0021853–C0042776
Cluster 3	
Angiotensins–Thirst	C0003018–C0039971
Thirst–Mouth	C0039971–C0922752
Mouth–Angiotensins	C0922752–C0003018
Cluster 4	
Sodium–Pregnancy	C0037473–C0032961
Pregnancy–Optic nerve	C0032961–C0029130
Optic nerve–Sodium	C0029130–C0037473

Table 9
Results for distances within the concept clusters

Conceptual distance measurement technique	Avg. within-cluster distance	STD
AVG Expert Score (MSH and SNMI)	6.7 (8.1)	2.7 (2.0)
AVG Expert Score for MSH	8.2 (8.1)	1.9 (2.0)
AVG Expert Score for SNMI	9.3 (9.2)	1.2 (1.2)

Table 10
Intra cluster distance results

Cluster pair	Avg. expert score	CDist (MSH)	CDist (SNMI)
Cluster1–Cluster2	3.36	5.66	6.66
Cluster1–Cluster3	9.67	7.44	8
Cluster1–Cluster4	9.67	8.11	8.11
Cluster2–Cluster3	9.67	7.77	8
Cluster2–Cluster4	9.67	8.44	8.11
Cluster3–Cluster4	10.59	8.88	8.33
Correlation W/avg-score	1.0	0.93	0.99

by the tight correlations. Notice that MSH results are missing two distances because concept Mouth, C0922752, is not a MSH term.

5.4. Using RB links to compute conceptual distance

We tried to compute CDist using RB (*related-broader*) links without any terminology restriction. We did not succeed because many of the distances could not be found, and the program was too slow to compute the rest. The results can be summarized in three observations:

1. When we used RB instead of PAR links, in many observations the number of linked concepts at each node increased by a factor of 8 or more, thus making the program extremely slow.
2. PAR links are not always included in the RB links, thus in practice there is no predictable relationship between distances computed using PAR and RB links.
3. The graph spanned by RB links shows many more disjoint sub-graphs than the graph spanned by PAR links. This makes it impossible to find paths between concepts for which paths are easily found in the PAR graph.

6. Discussion

The conceptual distance CDist, which is based on the shortest path between concepts along PAR links within a terminology source, shows potential as a conceptual similarity metric for use between concepts or concept clusters. The correlations obtained, although not impressive for an accurate mathematical model, show promise given the limited scope of the experiment. The potential benefits of a conceptual metric underline the advantages of the terminology desiderata that concepts may belong to multiple hierarchies, but they must be complete, and acyclic. In those conditions paths can be found and minimized among competing alternatives. Using CDist could potentially find similar concepts using independently developed connecting paths; and as we have shown, the metric can be used within selected terminology sources, or using combinations of them.

Further research is necessary to ascertain whether systematic errors such as underestimation of conceptual distance between general concepts are in fact a drawback of the approach, and to investigate possible solutions such as weighting the depth of the most specific common ancestor into the distance. Another area of research is the impact of flaws in the PAR hierarchies on CDist values. Although we have proposed to avoid topological errors such as cycles, dealing with such inconsistencies will require a more robust approach especially to avoid important errors in the CDist values.

We have shown that by comparing CDist values from different terminologies, it may be possible to identify important differences, and possibly inconsistencies, in graphs spanned by PAR links (e.g., by comparing distances in MSH, potentially the best performer, against other terminologies). One consequence of such graph inconsistencies/differences would be cycles in the joint graphs, or inability to find connecting paths in one of the terminologies. However, normal differences in graph density leading to marked differences in distance values (e.g., linear or non-linear scaling) across terminologies need to be further investigated.

Although it has been reported before that the shortest path along RB links in MeSH is a useful distance metric, in this work we have investigated the shortest PAR-link path, because of its practical applications within terminology sources and the flexibility to combine those sources (e.g., if a path between two concepts cannot be found in one terminology, it may be possible to find it in another one or by using joint terminologies). Our results showed that using unconstrained RB links proved too computationally expensive, and unreliable to find connecting paths. Moreover, using RB links constrained to a specific terminology, e.g., MSH, showed no performance improvement over the PAR links in our tests.

We also investigated distance between concept clusters (intra-cluster distance). After using a simple-minded approach, i.e., average pair wise distance, the results look encouraging. However, more research and experimental work (e.g., using document retrieval cases) is clearly required. Particularly, it would be necessary to incorporate principles such as semantic cluster structure, to give more relevance to more similar concepts across

clusters, and using semantic type information to decide how to weight other distances.

Using expert scores proved useful to test the performance of CDist. Although we did not feel the small number of experts substantially affected the results, in order to improve the experimental method it would be advisable to either involve more experts to reduce variance and enable full more formal statistical analysis, or resort to a panel-based scoring by a group of 3-5 top experts. The use of cognitive principles such as the relation between scoring time and similarity could be also incorporated to improve the robustness of the expert scores. A technique proposed in [16], which uses the principle that more dissimilar concepts take longer to score appears promising for future research.

Within the reduced scope of our experiment, it was observed that similar concepts clustered at the same distance from other unrelated concepts. This suggests that the triangle inequality property may be useful, as the distance between two concepts equally separated from a third one can be found from a separate estimate by the experts and compared against their CDist value, which must be $\geqslant 0$ and $\leqslant 2$ times the CDist to the third concept.

Regarding implementation, the CDist metric has been shown feasible through the use of Unix Shell scripts, which operate on the UMLS sources (MRSO and MRREL files). This is an advantage over more elaborate methods that would require porting the UMLS sources to a higher level representation and using recursive-programming techniques. The code has been written without any assumption about completeness or consistency of the hierarchies, thus it checks and avoids errors such as concepts that appear to be parent of themselves (e.g., C0178301|PAR|C0178301||ICD9CM|ICD9CM||), or circularities in the hierarchy.

Acknowledgments

Drs. James Cimino (JC), Michael Cantor (MC), and King Wah Fung (KW) from the Department of Biomedical Informatics, Columbia University, contributed the expert scores.

References

[1] Humpreys B, Lindberg D, Schoolman H, Barnett G. The Unified Medical Language: An Informatics Research Collaboration. JAMIA 1998;5(1):1–11.
[2] Aronson A. Meta-map: mapping text to the UMLS metathesaurus, Semantic Knowledge Representation Research Information Project. Available from: http://skr.nlm.nih.gov/papers.
[3] Saphire International '98 Web Site. Available from: http://www.ohsu.edu/cliniweb/saphint.
[4] Rada R, Mhashi M, Barlow J. Hierarchical semantic nets support retrieving and generating hypertext. Inform Decis Technol 1990;16(2):117–36.
[5] Tudhope D, Taylor C. Navigation Via Similarity: Automatic Linking Based on Semantic Closeness. Inform Process Manage 1997;33(2):233–42.
[6] Campbell J, Carpenter P, Sneiderman C, Cohn S, Chute C, Warren J. Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity. JAMIA 1997;4(3):238–51.
[7] Cimino J. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inform Med 1998;37:394–403.
[8] Lindberg D, Humpreys B, McCray A. The Unified Medical Language System. Methods Inform Med 1993;32:281–91.
[9] Rada R, Mili H, Bicknell E, Blettner M. Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybernet 1989;19(1):17–30, Jan/Feb.
[10] Tulving E. Episodic and semantic memory. In: Tulving E, Donaldson W, editors. Organization of memory. New York: Academic Press; 1972. p. 381–403.
[11] Tversky A. Features of similarity. Psychol Rev 1977;84:327–52.
[12] Kazman R, Al-Halimi R, Hunt W, Mantei M. Four paradigms for indexing video conferences. IEEE Multimedia Spring 1996:63–73.
[13] Aalbersberg IJ. A document retrieval model based on term frequency ranks. In: Proc. ACM SIGIR Conf. On Res. And Dev. in Information Retrieval, 1994, Dublin, Ireland. p. 163–72.
[14] Caviedes JE. An analogical reasoning engine for heuristic knowledge bases. In: Proceedings of the European Workshop on Case-Based Reasoning, EWCBR-93, University of Kaiserslautern, Germany, November 1993. p. 97–102.
[15] WordNet: a lexical database for the english language. Available from: http://www.cogsci.princeton.edu/~wn.
[16] Rubinsten O, Henik A. Is an ant larger than a lion. Acta Psychol 2002;111(1):141–54.