Latent Semantic Indexing of Medical Diagnoses Using UMLS Semantic Structures

C. G. Chute, M.D., Dr.P.H.†
Y. Yang, Ph.D.†
D. A. Evans, Ph.D.‡

†Section of Medical Information Resources, Mayo Clinic, Rochester, MN
‡Laboratory for Computational Linguistics, Carnegie Mellon University, Pittsburgh, PA

Then the little Hiawatha
learned of every bird its language,
learned their names and all their secrets...
--Longfellow

The relational files within the UMLS Metathesaurus contain rich semantic associations to main concepts. We invoked the technique of Latent Semantic Indexing to generate information matrices based on these relationships and created "semantic vectors" using singlular value decomposition. Evaluations were made on the complete set and subsets of Metathesaurus main concepts with the semantic type "Disease or Syndrome." Real number matrices were created with main concepts, lexical variants, synonyms, and associated expressions. Ancestors, children, siblings, and related terms were added to alternative matrices, preserving the hierarchical direction of the relation as the imaginary component of a complex number. Preliminary evaluation suggests that this technique is robust. A major advantage is the exploitation of semantic features which derive from a statistical decomposition of UMLS structures, possibly reducing dependence on the tedious construction of semantic frames by humans.

The goal of processing medical natural language for semiautomated classification remains elusive, although many efforts are promising[1]. The National Library of Medicine's Unified Medical Language System (UMLS) has potential in medical record classification[2]. Considerable research on the use of semantic networks to represent medical concept data exists[3][4][5], but their construction can be tedious.

Latent Semantic Indexing (LSI) is an Information Retrieval technique which takes advantage of term context and frequency to approximate a semantic basis for inquiry and retrieval[6]. Developed originally to classify and retrieve text documents independent of keywords or thesauri, its application to structured medical thesauri has been demonstrated[7]. We sought to evaluate the UMLS Metathesaurus as a structured thesaurus for the classification and inquiry of medical diagnoses using this technique.

Methods

Overview

LSI is premised on the construction of an information matrix wherein documents comprise the columns, and the words or terms make up the rows. This idea is modified so that UMLS main concepts constitute the columns, and words or terms associated with the concept are the rows. This matrix is then factored into principal components by singular value decomposition (SVD)[8]. A limited number (N) of factors and their related vectors are retained, which can be regarded as dimensions of a compressed, N dimensional semantic space. The truncated matrix deriving from the SVD can be used either to map inquiry phrases or to classify diagnostic text by reducing these strings to N-length vectors of factor weights unique to that string [6]. These vectors are then projected into the decomposed concept matrix, and the cosine of the vector in N-space is maximized against the concepts.

UMLS Inputs

The September, 1990, experimental CD-ROM version of the UMLS Metathesaurus (Meta-1) formed the input for these evaluations, (`mrxx`) notations below refer to files in the `metar` directory. Three concept lists were fashioned to evaluate the LSI technique in a preliminary way. The first we refer to as *Tiny-Input*, was composed of the following sample of Meta-1 main concepts: cerebral infarction; cholecystitis; cholelithiasis; coin lesion, pulmonary; coronary arteriosclerosis; kidney failure, acute; kidney tubular necrosis, acute; lung neoplasms; myocardial infarction; stroke. This was created to enable the presentation of a near complete example in this paper. The second is labeled *Midi-Input*

0195-4210/91/$5.00 © 1992 AMIA, Inc.

and consists of 101 concept strings selected by the authors as broadly representative of medical concepts. The third, called *Maxi-Input*, subsumes all 2,580 main concepts (mrmc) with a semantic type (mrsty) of "Disease or Syndrome." All examples were run with three inquiry lines:

carcinoma of the lung
myocardial infarction
cerebral ischemia.

For each input file, two classes of matrices were constructed: real and complex. Real matrices employed the main concepts themselves (10 for *Tiny-Input*, 101 for *Midi-Input*, and 2,580 for *Maxi-Input*), and their relationally-linked synonyms (mrsy), lexical variants (mrlv), and associated expressions (mratx). Complex matrices added UMLS entries that bore a hierarchical relationship to the main concepts in each input file, while preserving the direction of this hierarchy (parent vs. child) in an imaginary component of a complex number. Broader than entries (mranc, mrrrt type "Broader") contributed negative imaginary weights, narrower entries (mrchd, mrrrt type "Narrower") were positive, and peer concepts (mrsib, mrrrt type "Other") were neutral.

Canonicalization of Terms

Words and phrases can be fraught with number, case, tense, adverbial, or other inflections contributing to lexical variation. While explicit lexical variants are contained in the Metathesaurus, no enumeration can anticipate the combinational explosion of order and inflections encountered in every day applications. Our earlier invocation of lexical stemmers[2] partially ameliorated this difficulty, but occasionally introduced truncated strings in place of valid word roots. A word-oriented lexicon, with a sparing amount of potentially ambiguous multi-word terms, in which the canonical form of each word could be posted appeared necessary. Additionally, robust algorithms for reducing inflections not listed to an acceptable root must be coordinated with that lexicon. One of us (DAE) has been developing the morph tool within the CLARIT project[9] for nearly a decade, which was modified for our LSI evaluations.

We extensively edited a morph lexicon explicitly for this project. All adjective, adverbs, verbs, and inflectional variants were reduced to the singular noun form. A small library of near synonyms were then applied to indicate a preferred word root for similar words (e.g., tumor, neoplasm, carcinoma, and malignancy all reduce to cancer).

Matrix Construction

For each of the three input files, the main concepts and their relationally linked entries were canonicalized using our synonym-mapped morph lexicon. The resulting canonical words comprised matrix rows, while the columns were defined by the main concepts corresponding to the input file. Real number matrices were populated by 1s in cells at the intersection of a canonical word and a concept (i.e., that word was used in the main concept, synonym, lexical variant, or associated term); all other cells (most of them) were zero.

Complex matrix construction was similar, but added only 0.5 to the real number component if it derived from a hierarchical relation. The imaginary component is a weighted average of the entries that contributed to the cell, broader terms were -1, narrower were +1, all other terms 0; the real number component (1.0 or 0.5) served as the weight.

Numerical Solution

The singular value decomposition of the real matrix *Maxi-Input* was solved by a version of LINPACK SSVDC[10] specifically modified to converge on a matrix of this size; it took 1295 seconds on a Cray-Y/MP8 and somewhat over three weeks on a SUN SPARC 1+. The smaller, real and complex matrices were decomposed on a SUN SPARC 1+ using LINPACK's SSVDC and CSVDC in 1 to 500 seconds of CPU. We truncated the solution space (collapsed the factors) to N=500 for *Maxi-Input*, N=50 for *Midi-Input*, and N=5 for *Tiny-Input*.

Classification and Inquiry Strings

The method outlined by Deerwester[6] was employed to create matching vectors from phrases and to project these into the matrix solution spaces. Given an input matrix X, the process of the SVD, $X=U_o\Sigma_oV_o^T$, yields three outputs: the singular values Σ_o, U_o (concept x word dimensions), and V_o (a square matrix, with rank equal to the smaller dimension of the input matrix, typically the number of concepts). Each array has one dimension truncated to N, which compresses the semantic space and reduces the computational demand of applying the decomposition. The truncation yields an approximation $X \approx \hat{X} = U\Sigma V^T$ where the dimensions of U are number of canonical words x N, those in V are the number of concepts in input file x N, and Σ includes the N most significant singular values. U, V and Σ are used in mapping an inquiry phrase to the term space. Inquiry phrases or medical text to be classified are canonicalized using morph and represented as a vector X_q similar to a column of X. X_q is transformed into the new vector space: $V_q=X_q^TU\Sigma^{-1}$. Finally, the distance between the inquiry and each term is measured by the cosine θ:

$$\cos\theta = \frac{(V_q \cdot V_i)}{\|V_q\| \times \|V_i\|}$$

where V_q is the transformed inquiry vector, V_i is a column of the V matrix; "·" = dot product, $\|\ \|$ = Euclidean norm.

Results

Figure 1 represents the first 20 canonical word stems which derived from the matching for *Tiny-Input*, 153 stems were created. The upper slice of the initial information matrix generated for this 153x10 cell complex input for decomposition is depicted in Figure 2. This matrix is striking for its sparseness, a finding that is only exaggerated for the larger matrices.

Figure 1
First 20 Term Stems Derived from *Tiny-Input*

abscess
arteriovenous
acute
adult
alveolar
aneurysm
angina
anoxia
anuria
arteria
arteriosclerosis
artery
aspiration
atelectasis
atresia
bile
bleed
bronchial
calculus
carcinoma

Figure 2
Partial Information Matrix for *Tiny-Input*

MC->:	Cerebral Infarction	Cholecys- titis	Choleli- thiasis	Coin Lesion, Pulmonary	Coronary Arterio-	Kidney Failure,	Kidney Tubular	Lung Neoplasms	Myocardial Infarction	Stroke
abscess	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)
acute	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(1.0, 0.1)	(1.0,-0.2)	(0.0, 0.0)	(0.5, 1.0)	(0.5, 0.0)
adult	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)
alveolar	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)
aneurysm	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)
angina	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)
anoxia	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)
anuria	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)
arteria	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5,-1.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)
arteriosclerosis	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(1.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)
arteriovenous	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)
artery	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)
aspiration	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)
atelectasis	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)
atresia	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)
bile	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.5)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)
bleed	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)
bronchial	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 1.0)	(0.0, 0.0)	(0.0, 0.0)
calculus	(0.0, 0.0)	(0.0, 0.0)	(1.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)
carcinoma	(0.0, 0.0)	(0.5, 0.0)	(0.5, 0.0)	(0.5,-0.6)	(0.0, 0.0)	(0.5, 0.0)	(0.0, 0.0)	(1.0, 0.0)	(0.0, 0.0)	(0.0, 0.0)

Note complex number pairs and very sparse information content; a single complex number is in parenthesis, the comma separates the real and imaginary components in that order.

The five largest complex cosine values for *Tiny-Input* are shown in Figure 3. The magnitude of the imaginary component, representing the deviation from the hierarchy, is very sensitive to program parameters (number of dimensions chosen, weighting values, etc.). We have not yet established optimum values for these parameters, and defer comment on their interpretation until we are confident in our parameter selection. Parameters for this printout include a truncation value (N) of 5, and an axis deviation weight of 0.5.

The *Midi-Input* dataset provided more interesting results. Among the 101 concepts, a total of 258 word stems were generated in the real number decomposition, while 1,197 were identified for the complex run. The complex run generated more stems because ancestor and child terms were identified. The real number (non-hierarchical) output for the five closest matching concepts (among the 101) are shown in Figure 4.

The largest file, *Maxi-Input*, contained a total of 2,580 concepts which yielded 3,803 word stems in the non-hierarchical (real number decomposition) evaluation, while 5,434 stems derived from the complex analysis. The SVD algorithm required 1300 seconds of CPU on a Cray- Y/MP8 computer, 6,000 on a Cray-2, 32,000 on an i860 SKYstation vector accelerator, and three weeks (wall clock time) on a SPARC 1+; it consumed over 65 Mb of virtual memory. No attempt was made to solve the complex matrix. The output for the evaluation inquiry statements appears in Figure 5.

Discussion

While the computational demands to solve the initial decomposition can be formidable, once achieved for a static problem (e.g. an "edition" of the UMLS), it need not be recomputed. The machine resources needed for classification or inquiry of patient data are in fact quite modest.

The greatest power of the methodology lies in its invocation of semantic dimensions that are derived from computable data structures by statistical methodologies.

Thus, major revisions in the source databases do not present an obstacle to maintaining consistent semantic representations without large amounts of tedious human effort (albeit the validity and utility of these semantic structures have yet to undergo exhaustive scrutiny).

The resulting semantic structures are only as complete as the relational input data. "True" semantic relations between terms that are not included in the information matrix computation or are not represented in the UMLS, will not be recognized. This raises the issue of attempting to edit the UMLS to represent these broader content relations, a view that was judged inappropriate during the initial Meta-1 editing process[11]. It is addressed in part by our `morph` lexicon, but effectively only at the word level.

Figure 3
Complete LSI Process Output for *Tiny-Input*

Input Phrase	Matched Phrase	Cosine Deviation
carcinoma of the lung	Lung Neoplasms	(0.99,-0.01)
	Coin Lesion, Pulmonary	(0.99,0.01)
	Cholecystitis	(0.42,-0.10)
	Cholelithiasis	(0.18,-0.15)
	Kidney Failure, Acute	(0.17,-0.11)
cerebral ischemia	Stroke	(0.98,-0.02)
	Cerebral Infarction	(0.91,0.00)
	Coronary Arterio-sclerosis	(0.18,-0.06)
	Cholecystitis	(0.03,-0.03)
	Myocardial Infarction	(0.13,-.11)
myocardial infarction	Myocardial Infarction	(0.93,-0.03)
	Coronary Arterio-sclerosis	(0.71,-0.06)
	Cholecystitis	(0.26,-0.17)
	Cerebral Infarction	(-0.00,-0.02)
	Coin Lesion, Pulmonary	(-0.02,-0.01)

Note that the cosine of the concept deviations are complex values.

Many natural experiments and evaluations present themselves with this innovative technique. Vast combinations of parameter settings, lexical input processing, cosine value manipulation, and scoring remain to be tried and tested. The relatively impressive performance of the technique with "wild guess" estimates of these values suggests that the method may be very robust.

We conclude that LSI may function remarkably well as a mechanism for classifying and retrieving patient record text data. It invokes semantic structures which derive from UMLS constructs. Much more evaluation remains to be done, but preliminary efforts are promising.

Figure 4
Top Level Matches for Real Number Decomposition of *Midi-Input*

Input Phrase	Matched Phrase	Cosine Deviation
carcinoma of the lung	Lung Neoplasms	1.00
	Coin Lesion, Pulmonary	0.93
	Chronic Obstructive Pulmonary Disease	0.55
	Anemia	0.27
	Dermatitis, Contact	0.09
cerebral ischemia	Cerebral Ischemia, Transient	0.90
	Cerebral Infarction	0.53
	Angina, Unstable	0.39
	Myocardial Infarction	0.22
	Abdomen, Acute	0.10
myocardial infarction	Myocardial Infarction	0.81
	Angina, Unstable	0.67
	Cerebral Infarction	0.54
	Heart Block	0.29
	Bronchitis	0.19

Top matches from intermediate-sized information matrix (101 concepts).

Figure 5
Top Level Matches for Real Number Decomposition of *Maxi-Input*

Input Phrase	Matched Phrase	Cosine Deviation
carcinoma of the lung	CARCINOMA OF LUNG	1.00
	Carcinoma, Non-Small Cell Lung	1.00
	Lung Neoplasms	1.00
	Pleural Neoplasms	1.00
	Bronchial Neoplasms	0.85
cerebral ischemia	Cerebral Ischemia	0.80
	Cerebral Infarction	0.64
	Encephalomalacia	0.64
	Brain Damage, Chronic	0.60
	Cerebral Ischemia, Transient	0.60
myocardial infarction	Myocardial Infarction	0.78
	Myocardial Reperfusion Injury	0.65
	Myocardial Diseases	0.59
	Angina, Unstable	0.48
	Cerebral Infarction	0.25

Top matches from all 2,580 main concepts in UMLS with semantic type "Disease or Syndrome."

Acknowledgements

We are grateful for the programming support of Geoff Atkin, the nosologic editing of Donna Irhke, and to Karen Elias for manuscript preparation. Cray Research, Eagan Park, MN, provided time on their Y/MP8 demonstration machine. This work is supported in part by a grant from NIH (AR30582).

References

1. Chute CG, Côté RA. Computerized natural medical language processing for knowledge representation: Overview of IMIA Working Group Conference, Geneva, September, 1988. In: B. Barber, D. Cao, D. Qin, G. Wagner (eds.) *Proceedings of the Sixth Conference on Medical Informatics, Beijing, China, October 16-20, 1989 and Singapore, Republic of Singapore, December 11-15, 1989*, 1989.
2. Chute CG, Yang Y, Tuttle MS, Sherertz DD, Olson NE, Erlbaum MS. *A preliminary evaluation of the UMLS Metathesaurus for patient record classification. Proceedings of the Fourteenth Annual*Symposium on Computer Applications in Medical Care, 1990;161-165.
3. Barr CE, Komorowski HJ, Pattison-Gordon E, Greenes RA. Conceptual modeling for the Unified Medical Language System. In: RA Greenes (ed). *Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care* 1988;148-151.
4. Evans DA. Pragmatically-structured, lexical-semantic knowledge bases for Unified Medical Language Systems. In: R. A. Greenes (ed). *Proceedings of the Twelfth Annual Symposium on Computer Applications in Medical Care* 1988;169-173.
5. Hersh WR, Greenes RA. SAPHIRE--An information retrieval system featuring concept matching, automatic indexing, probablistic retrieval, and hierarchical relationships. *Computers and Biomedical Research* 1990;23(5):410-425.
6. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science* 1990;41(6):391-407.
7. Evans DA, Handerson SK, Monarch IA, Pereiro J, Delon L, Hersh WR. Mapping vocabularies using "Latent Semantics." *Technical Report No. CMU-LCL-91-1*, Pittsburgh, PA: Computational Linguistics Laboratory, Carnegie Mellon University, 1991.
8. Golub CH, Van Loan CF. *Matrix Computations*. Oxford: North Oxford Academic, 1983.
9. Evans DA, Ginther-Webster K, Hart M, Lefferts RG, Monarch IA. Automatic indexing using selective NLP and First-Order Thesauri. *RIAO '91*, April 2-5, 1991, Autonoma University of Barcelona, Barcelona, Spain, pp. 624-644.
10. Dongarra JJ, Moler CB, Bunch JR, Stewart GW. *LINPACK Users' Guide*. Philadelphia, PA: SIAM, 1979.
11. Sperzel D, Erlbaum M, Fuller L, Sherertz D, Olson N, Schuyler P, Hole W, Savage A, Passarelli P, Tuttle M. Editing the UMLS Metathesaurus: Review and enhancement of a computed knowledge source. In: RA Miller (ed) *Proceedings of the Fourteenth Annual Symposium on Computers in Medical Care* 1990;136-140.