Predication-based Semantic Indexing: Permutations as a Means to Encode Predications in Semantic Space

Trevor Cohen, MBChB, PhD[a], Roger W. Schvaneveldt, PhD[b], Thomas C. Rindflesch, PhD[c]
[a]Center for Decision Making and Cognition, Department of Biomedical Informatics, Arizona State University, Phoenix Arizona
[b]Department of Applied Psychology, Arizona State University
[c]National Library of Medicine, Bethesda, Maryland

Abstract
Corpus-derived distributional models of semantic distance between terms have proved useful in a number of applications. For both theoretical and practical reasons, it is desirable to extend these models to encode discrete concepts and the ways in which they are related to one another. In this paper, we present a novel vector space model that encodes semantic predications derived from MEDLINE by the SemRep system into a compact spatial representation. The associations captured by this method are of a different and complementary nature to those derived by traditional vector space models, and the encoding of predication types presents new possibilities for knowledge discovery and information retrieval.

Introduction
The biomedical literature contains vast amounts of knowledge that could inform our understanding of human health and disease. Much of this literature is available as electronic text, presenting an opportunity for the development of automated methods to extract and encode knowledge in computer-interpretable form. Distributional models of language are able to extract meaningful estimates of the semantic relatedness between terms from unannotated free text. These models have proved useful in a variety of biomedical applications (for a review see (1)), and include recent variants that scale comfortably to large biomedical corpora such as the MEDLINE corpus of abstracts (2).

However, the semantic relatedness estimated by most distributional models is of a general nature. These models do not encode the type of relationship that exists between terms, which limits their ability to support logical inference. Furthermore, while distributional models such as Latent Semantic Analysis (LSA) simulate human performance in many cognitive tasks (3), they do not represent the object-relation-object triplets (or propositions) that are considered to be the atomic unit of thought in cognitive theories of comprehension (4). In this paper we address these issues by defining Predication-based Semantic Indexing (PSI), a novel distributional model of language that encodes semantic predications derived from MEDLINE by the SemRep system (5) into a compact vector space representation. Associations captured by PSI complement those captured by existing models, and present new possibilities for knowledge discovery and information retrieval.

Background
Many existing distributional models draw estimates of semantic relatedness from co-occurrence statistics within a defined context such as a sliding window or an entire document (1). Recent models (reviewed in (6)) instead define as a context a grammatical relationship produced by a parser, but do not encode the nature of this relationship in a retrievable manner. Distributional models that encode word order using either convolution products (7) or permutation of sparse random vectors (8) transform vectors representing terms into new representations close-to-orthogonal to the original vectors. Consequently there is minimal overlap in the information they carry, and additional information related to term position can be encoded. These transformations are reversible, to facilitate retrieval of this information.

PSI is based on Sahlgren *et al*'s model which uses permutations as a means to encode word order information (8), which in turn is a variant of the Random Indexing (RI) model (9). Sahlgren *et al*'s approach provides a simple and elegant solution to the problem of reversibly transforming term vectors using permutations of the sparse random vectors which form the basis of RI. The approach is derived from sliding-window (or term-term) RI, derives vector representations for terms from their co-occurrence with other terms in a sliding window moved through the text. While the sliding window approach is well-established in distributional semantics, established methods either use the full term-term space or reduce its dimensionality with the computationally demanding Singular Value Decomposition (SVD). RI is able to achieve this dimension reduction step at a fraction of the cost of SVD by constructing semantic vectors for each term on-the-fly, without the need for a term-by-term matrix. Each term in the text corpus is assigned an elemental vector of dimensionality *d* (usually in the order of 1000), the dimensionality of a reduced-dimensional *semantic space* within which the relatedness of terms will be measured. Elemental vectors are sparse: they contain mostly zeros, with in the order of 10 non-zero values of either +1 or -1. As there are many possible permutations of these few non-zero values, elemental vectors tend to be close-to-orthogonal to one another: their relatedness as measured with the commonly used cosine metric tends towards zero. This approximates a full term-by-term matrix, but rather than assigning an orthogonal dimension to each term, RI assigns a near-orthogonal

reduced-dimensional elemental vector. To encode additional information to do with word order, the elemental vector for a given term is permuted to produce a new vector, almost orthogonal to the vector from which it originated. Consider vectors below:

V1: [1 0 0 0 0 1 0 0 0 0 0 -1 0 0 0] *V2*: [0 1 0 0 0 0 1 0 0 0 0 0 -1 0 0]

These vectors are orthogonal to one another: as there is no common non-zero dimension between them, their cosine (or normalized dot-product) will be zero. V2 was derived from V1 by moving every value one position to the right, and conversely this transformation can be reversed by moving every value in V2 one position to the left. This simple procedure is used by Sahlgren *et al* to encode word-order information into a term-term based semantic space. The semantic vector for each term consists of the normalized linear sum of the permuted elemental vector for every term with which it co-occurs, with permutation encoding the relative position of each term in the sliding window. The reversible nature of this transformation facilitates order-based retrieval. For example, a rotation one position to the right of all elements of the elemental vector for a term can be used to generate a vector with high similarity to terms occurring one space to the left of it. *Table I* provides some examples of order-based retrieval in a permutation-based space derived from the MEDLINE corpus of abstracts using the Semantic Vectors package (10).

? cancer	streptococcus ?	? cough
.81:breast	.71:pyogenes	.89:whooping
.78:colorectal	.71:agalactiae	.48:nonproductive
.74:prostate	.69:pyogens	.47:hacking
.69:antiprostate	.65:milleri	.44:brassy
.67:antibreast	.62:acidominimus	.42:barking

Table I*: Order-based retrieval from MEDLINE. The "?" denotes the relative position of the target term.*

In this paper, we adapt Sahlgren *et al*'s method of encoding word order information into a vector space to encode semantic predications produced by the SemRep system (5), (11). SemRep combines general linguistic processing, a shallow categorical parser and underspecified dependency grammar, with domain-specific knowledge resources: mappings from free text to the UMLS accomplished by the MetaMap software (12), the UMLS metathesaurus and semantic network (13) and the Specialist lexicon and lexical tools (14). SemRep uses these techniques to extract semantic predications, from titles and abstracts in the MEDLINE database, as shown in this example drawn from (5). Given the excerpt "*... anti-inflammatory drugs that have clinical efficacy in the management of asthma,....*", SemRep extracts the following semantic predication between UMLS concepts:

"Anti-Inflammatory Agents TREATS Asthma"

We present in this paper a description of the theoretical and methodological basis of PSI, and include examples of the sorts of information the model encodes and retrieves discussed in context of possible applications.

Methods

We derived a PSI space from a database of semantic predications extracted by SemRep from MEDLINE citations dated between 2003 and September 9th 2008. 13,562,350 predications were extracted from 2,634,406 citations by SemRep. Of these, predications involving negation (such as "DOES NOT TREAT") are excluded, leaving 13,380,712 predications which are encoded into the PSI space**.** We encode this predication information using permutation-based RI. Rather than assigning elemental vectors to each term, we assign sparse elemental vectors (d=500) to each UMLS concept contained in the predications database. We then assign a unique number to each of the included predication types (such as "TREATS"). We create semantic vectors (d=500) for each UMLS concept in the database. Each time a given UMLS concept occurs in a predication, we add to its semantic vector the elemental vector of the other concept in the predication, permuted according to the predication type. For example, in the predication "Isoniazid TREATS Tuberculosis" we would add the elemental vector for Tuberculosis (TB) to the semantic vector for Isoniazid (INH) but rotate every element in this elemental vector 39 (the number assigned to the predicate "TREATS") steps to the left. Conversely, we would add to the semantic vector for TB the elemental vector for INH rotated 39 steps to the right. In this way we can encode the predication connecting these concepts.

We also construct a general distributional model of the UMLS concepts in the database of predications using the Reflective Random Indexing (RRI) model (15), by creating document vectors for each unique PubMed ID in the database. Document vectors are created based on the terms contained in these citations: elemental vectors are assigned to each term, and document vectors are constructed as the normalized linear sum of the elemental vector for each term they contain. Rather than using raw term frequency, we employ the log-entropy weighting scheme, shown to enhance document representations in several applications (3). A vector for each concept is constructed as the frequency-weighted normalized linear sum of the vector for each document it occurs in.

PSI requires a modification of the conventional nearest neighbor approach, as we are interested in the strongest association between concepts across all predications. In the modified semantic network used by SemRep (16), there are 40 permitted predications between concepts when negations (e.g. exercise DOES NOT TREAT hiv) are excluded. Semantic distance in PSI is measured by extracting all permutations of a concept, and comparing the second concept to these to find the predication with

the strongest association. For elemental vectors, we employ a sparse representation used in our previous work (2) which represents the dimension and sign of each of the 20 non-zero values. This allows for rapid generation of all possible permutations by augmenting the value that represents the index of each non-zero value. To further speed up this process in the EpiphaNet example *(Figure 1)*, we extract the 500 nearest neighbors to a cue concept from the general distributional space (this should subsume the predication-based space: every concept in a predication must co-occur in a citation with the other concept concerned). We then perform predication-based nearest-neighbor search on these neighbors only. As it is possible to search either using elemental vectors as cues to retrieve semantic vectors or vice-versa, for the quantitative evaluations we assess associations in both directions to ensure accessing the strongest association.

Results and Discussion

Predication-based retrieval

In a manner analogous to the order-based retrieval illustrated previously, it is possible to perform predication-based retrieval using permutations to determine which UMLS concept the model has encoded with strong association to another concept in a particular predication relationship. *Table II* illustrates predication-based retrieval. For example, the query "? TREATS Asthma" retrieves concepts for asthma treatments (sb-240563, also known as Mepolizumab, has recently been shown to reduce exacerbations in asthma (17)) .

? TREATS Asthma	Metronidazole TREATS ?
1:cetirizine-pseudoephedrine	0.57: chronic intestinal amebiasis
1: norisodrine	0.36 : urogenital trichomonas nos
1: alvesco	0.35: erythema annulare centrifugum
1: salmeterol+fluticasone propionate	0.33: vaginalis
1: sb-240563	0.27: endocervicitis, unspecified

***Table II**: Predication-based retrieval with cosine associations between query and target concepts.*

Interestingly, the top ranked results are not necessarily the concepts that occur most frequently in this predication relationship. Rather, these results reflect the extent to which this relationship defines a particular concept, as the model represents concepts in terms of the predications in which they occur in an extensional manner. Concepts occurring exclusively in a particular predication with another concept are likely to rank highly in predication-based retrieval. As this is not ideal for many purposes, our future work will explore variants of PSI that select for frequency rather than exclusivity.

Predication-based Nearest Neighbor Search

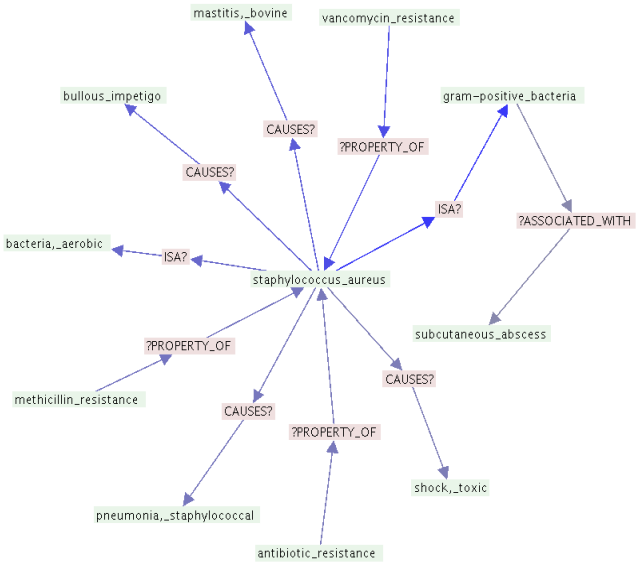

***Figure I**: EpiphaNet for "staphylococcus aureus"*

It is possible to rapidly characterize a particular concept for exploratory purposes by first finding the *k*-nearest neighbors in a general associative space, and searching amongst these for the best predications using PSI. *Figure I* illustrates the nearest predication-based neighbors of the concept "staphylococcus_aureus" which we have extracted and visualized with the EpiphaNet software we have developed for this purpose. EpiphaNet is based on the Prefuse visualization library (18) and as in our previous work (2) uses Pathfinder network scaling (19) to reveal the most significant associative links within a network of near neighbors. By reversing the encoding process used in PSI, we are able to retrieve both the type and direction of the predication relationship linking these concepts. This measure of semantic distance is different in nature to those used in prior distributional models. Rather than conflating many types of association into a single metric, this estimate is based on the strongest typed association between these concepts across all predications. Similar to the way in which existing distributional models extract compact vector-based term representations from large corpora, the PSI model produces a compact representation for all UMLS concepts in the 8.8GB database of semantic predications. The set of semantic vectors used for the PSI space used to generate *Figure I* occupies 300MB only, and stored elemental vectors occupy a fraction of this space due to the sparse representation employed. To further assess the extent to which predications are accurately encoded and retrieved, we extract at random 1000 concepts, and retrieve their 20 nearest predication-based neighbors. We consider neighbors with a cosine association above a threshold of the mean cue-to-neighbor association for these 1000 terms as "*retrieved*". Using the database of predications extracted by SemRep as a gold standard, we calculate the following metrics:

o Precision = *retrieved* and accurate / all *retrieved*
o Recall = predications *retrieved* / minimum(20, *up*)

where *up* denotes the number of unique predications for cue term in the database. Results are shown in *Table III*.

Dimensionality	500	1000	1500
Mean Precision	0.957	0.977	0.997
Mean Recall	0.603	0.643	0.658
Threshold cosine	0.320	0.279	0.265

Table III*: Results for 1000 randomly selected concepts.*

The model performs better for cue concepts with fewer unique predications: recall when only concepts with 20 or less unique predications are considered is 0.74, 0.8 and 0.8 at 500, 1000 and 1500 dimensions respectively, with precision at 0.95 and above. This suggests that vectors for concepts involved in many predication relationships acquire a spurious similarity to other vectors due to partial overlap between permuted elemental vectors. We anticipate this overlap would reduce as dimensionality increases. In practice we find that concepts such as "patient" that are involved in many unique predications tend to be uninformative. It is also possible to eliminate spurious neighbors by only considering terms that occur in a document with the cue term as retrieval candidates.

Implicit Encoding of Semantic Type
As illustrated by the results of the cosine-based nearest neighbor search in *Table IV,* the PSI space to some extent captures the semantic class of UMLS concepts.

asthma	amitryptiline
.98: sickle cell anemia .99 : heart septal defects, atrial .99: chronic childhood arthritis .98: diarrhea .98: constipation	.82: imipramine .78: nortriptyline .76: desipramine .75: clomipramine .65: amoxapine

Table IV: *Nearest-neighbor searches in PSI-space.*

The semantic vector for the disease "asthma" is similar to that for other diseases (and in this case, symptoms), just as "amitryptiline" retrieves other antidepressants through nearest neighbor search. This finding generalizes to a degree: amongst the ten-nearest neighbors of 1000 randomly selected terms, an average of 37% share a UMLS semantic type with the cue term. This is considerably higher than the result of approximately 5% obtained when the same evaluation is performed using either RI (9)or RRI (15) (all spaces at d=500), and varies across semantic types, with several semantic classes such as "plant" exhibiting in excess of 80% agreement between cue and neighbor. This is to be expected, as the extraction of predications by SemRep is constrained by the UMLS semantic type of the subject and object. However, further analysis of the interplay between assigned UMLS class and predication-based distributional similarity may be a useful way to reveal inconsistencies in the assignment of semantic class and/or the assignment of predications by SemRep.

Modeling Analogy
We find it is possible to model analogy within the PSI space by finding the predication that most strongly associates two terms and applying the rotation that corresponds to this predication to a third. While this work is presently at an early stage of development, it has produced some interesting results so far (*Table V)*.

Example	Cue	Retrieved
Tuberculosis is to Isoniazid as.....	Depressive disorder is to..	Lexapro
Tuberculosis is to Lung X-ray as...	Depressive disorder is to..	Psychiatric Interview and Evaluation

Table V*: Analogical reasoning in PSI-space.*

Application to Information Retrieval
Similarly to the way in which distributional models extract compact vector-based term representations from large corpora, the PSI model produces a compact representation of the predication relations captured by SemRep. The knowledge encoded in the PSI model could be used for information retrieval in several ways. One possibility would be to represent documents in terms of the predications contained therein, and allow users to search for documents containing concepts in a specific predication relationship with a search concept. We anticipate that once customized for this purpose, PSI will retrieve documents providing answers to clinical questions such as "what treats Tuberculosis" or "what causes Bullous Impetigo". Another possibility would be the use of the approach taken in *Figure I* to categorize documents according to the way in which they are related to a particular search concept. In our future work we will evaluate these approaches on standard test collections.

Application to Literature-based Knowledge Discovery
In our recent work (2),(20),(15) we have used general distributional models to identify potential discoveries by identifying pairs of concepts that are relatively close in the space but do not co-occur in any of the documents in the database used to generate the models. Although this method has proven to be effective in identifying interesting indirect connections, the interesting ones tend to occur along with others of little interest. In general, additional constraints are needed to narrow the possibilities. The predications resulting from the methods presented here offer a promising means to limit the indirect connections by selecting those with appropriate predication relationships. For example, when looking for new treatments for a disorder, concepts that serve as treatments should be given priority over concepts in other predications. With these methods, general word space

similarity can be elaborated into the greater specificity found in semantic network models (21).

Limitations and Future Work

This paper presents the theoretical and methodological basis for PSI, a novel distributional model that encodes predications produced by SemRep, and provides some illustrative examples and possible applications. Further analysis is needed to determine the model parameters that optimize performance in each of these tasks. We do not evaluate the performance of SemRep, as this has been evaluated elsewhere (5,16). In our future work we will explore applications of PSI to informatics problems, including information retrieval, knowledge discovery and biomedical question answering.

Conclusion

PSI is a novel distributional model that encodes predications produced by the SemRep system, providing a more specific measure of semantic similarity between concepts than is provided by existing distributional models, as well as the ability to retrieve the type of predication that most strongly associates two concepts. From a theoretical perspective, this is desirable as the unit of analysis in cognitive models is considered to be an object-relation-object triplet, not an individual term. From a practical point of view, the additional information encoded by PSI is likely to be of benefit for information retrieval and knowledge discovery purposes. In our future work we will evaluate the application of PSI to these and other informatics problems.

Acknowledgments

We would like to acknowledge Dominic Widdows, chief instigator of Semantic Vectors (10), some of which was adapted to this work, and Sahlgren, Holst and Kanerva for their remarkable contribution to the field.

References

1. Cohen T, Widdows D. Empirical distributional semantics: Methods and biomedical applications. JBI, 2009 Apr;42(2):390-405.
2. Cohen TA. Exploring MEDLINE Space with Random Indexing and Pathfinder Networks. AMIA Annu Symp Proc. 2008 ;126-30.
3. Landauer TK, Dumais ST. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psych. Review. 1997 ;104211-240.
4. Kintsch W. Comprehension : a paradigm for cognition. Cambridge, ; New York, NY: Cambridge University Press; 1998.
5. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. JBI 2003;36462-477
6. Pado S, Lapata M. Dependency-Based Construction of Semantic Space Models. Computational Linguistics. 2007 ;33161-199.
7. Jones MN, Mewhort DJK. Representing word meaning and order information in a composite holographic lexicon. Psych. Review. 2007 ;1141-37.
8. Sahlgren M, Holst A, Kanerva P. Permutations as a Means to Encode Order in Word Space. Proc. 30th Annual Meeting of the Cognitive Science Society (CogSci'08), July 23-26, Washington D.C.; 2008 ;
9. Kanerva P, Kristofersson J, Holst A. Random indexing of text samples for latent semantic analysis. Proc. of 22nd Annual Conference of the Cognitive Science Society. 2000 ;1036
10. Widdows D, Ferraro K. Semantic Vectors: A Scalable Open Source Package and Online Technology Management Application. Sixth International Conference on Language Resources and Evaluation (LREC 2008);
11. Rindflesch TC, Fiszman M, Libbus B. Semantic interpretation for the biomedical research literature. Medical informatics: Knowledge management and data mining in biomedicine. 2005 ;
12. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001 ;1717-21.
13. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Research. 2004; 32 (Database Issue):D267.
14. Browne AC, Divita G, Aronson AR, McCray AT. UMLS language and vocabulary tools. In: AMIA Annu Symp Proc. 2003. p. 798.
15. Cohen T, Schvaneveldt R, Widdows D. Reflective Random Indexing and Indirect Inference: A Scalable Method for the Discovery of Implicit Connections. (manuscript under review)
16. Ahlers CB, Fiszman M, Demner-Fushman D, Lang FM, Rindflesch TC. Extracting semantic predications from Medline citations for pharmacogenomics. Pac Symp Biocomput. 2007:209-20.
17. Haldar P, Brightling CE, Hargadon B, Gupta S, Monteiro W, Sousa A, et al. Mepolizumab and exacerbations of refractory eosinophilic asthma. N Engl J Med. 2009 Mar 5;360(10):973-84.
18. Heer J, Card SK, Landay JA. prefuse: a toolkit for interactive information visualization. Human Factors in Computing Systems. 2005 ;421-430.
19. Schvaneveldt RW. Pathfinder associative networks: studies in knowledge organization. Ablex Publishing Corp. Norwood, NJ, USA; 1990.
20. Schvaneveldt, RW, Cohen, TA. Abductive Reasoning and Similarity. In: In: Ifenthaler D, Seel NM, editor(s). Computer based diagnostics and systematic analysis of knowledge. Springer, NY;
21. Quillian MR. Semantic memory. Minsky, M., Ed. Semantic Information Processing. 1968 ;216-270.