Finding Schizophrenia's Prozac:
Emergent Relational Similarity in Predication Space

Trevor Cohen[1], Dominic Widdows[2] Roger Schvaneveldt[3], and Thomas C. Rindflesch[4]

[1] University of Texas Health Science Center at Houston
[2] Google, inc.
[3] Arizona State University
[4] National Library of Medicine

Abstract. In this paper, we investigate the ability of the Predication-based Semantic Indexing (PSI) approach, which incorporates both symbolic and distributional information, to support inference on the basis of structural similarity. For example, given a pair of related concepts prozac:depression, we attempt to identify concepts that relate to a third concept, such as schizophrenia in the same way. A novel PSI implementation based on Kanerva's Binary Spatter Code is developed, and evaluated on over 100,000 searches across 180,285 unique concepts and multiple typed relations. PSI is shown to retrieve with accuracy concepts on the basis of shared single and paired relations, given either a single strong example pair, or the superposition of a set of weaker examples. Search space size is identical for single and double relations, providing an efficient means to direct search across predicate paths for the purpose of literature-based discovery.

Keywords: Distributional Semantics, Vector Symbolic Architectures, Literature-based Discovery, Abductive Reasoning

1 Introduction

This paper presents new results that demonstrate ways in which high-dimensional vector representations can be used to model proportional analogies such as "prozac is to depression as what is to schizophrenia?" Our approach is based on our earlier "Logical Leaps" work [1], and Kanerva's work on hyperdimensional computing and analogical mapping [2] (both presented at Quantum Informatics, 2010). This approach depends upon being able to represent concepts as high-dimensional vectors, and relationships between concepts as mathematical operations on these vectors. Such operations include composition of vectors using product and superposition operations, and the selection of nearby pure concepts from a superposed or product state. The work is part of the family of generalized quantum methods currently being explored: basic concepts are analogous to pure states; superposition and product operations give rise to compound concepts analogous to mixed and entangled states; and the selection of a nearby known concept from a product state is analogous to quantization or quantum collapse. A notable departure from traditional quantum mechanics is our use of real and binary vectors, instead of complex vectors. This departure is not novel and is an oft-understated discrepancy of approaches: for many years the information retrieval and machine learning communities have used real-valued vectors; Kanerva's work uses binary-valued vectors as examples [2]; and traditional quantum mechanics almost exclusively used complex Hilbert spaces, as have emerging approaches to information retrieval [3] and distributional semantics [4]. We mention this at the outset as perhaps one of the key senses in which "generalized quantum" models should be thought of as generalizations, not applications, of quantum physics.

2 Background

The "Logical Leaps" approach is an extension of our previous work in the domain of literature-based discovery [5], in which we evaluated the ability of various scalable models of distributional semantics to generate *indirect inferences* [6], meaningful connections between terms that do not co-occur in any document in a given corpus. Connections of this sort are fundamental to Swanson's model of literature-based discovery [7], which emerged from the serendipitous discovery of a therapeutically useful [8] connection between Raynaud's Syndrome (reduced blood flow in the extremities) and fish oils. This connection was based on the bridging concept "blood viscosity": fish oil can decrease blood viscosity thus increasing blood flow. Swanson's method can be seen as an example of abductive reasoning, hypothesis generation as proposed by Peirce (see [9]), and provides the basis for several computer models that aim to facilitate discovery [10], [11]. As an alternative to stepwise exploration of the vast search space of possible bridging concepts and discoveries, distributional approaches such as Latent Semantic Analysis [6], Random Indexing (RI) [12] and others have been applied to infer meaningful indirect connections between terms without identifying a bridging concept [13], [14], [5]. In contrast to these approaches, which are based on general association strength, "Logical Leaps" are derived from a vector space in which both the target and the type of a relation to a concept are encoded into its vector representation. This has been achieved using Predication-based Semantic Indexing (PSI) [15], a variant of RI that uses permutation of sparse random vectors to encode relationships (such as TREATS) between concepts into a high-dimensional vector space. In this paper, we attempt to direct searches in PSI space by specifying predicate paths using a pair of example concepts. We achieve this end with an alternative implementation of PSI based on Kanerva's Binary Spatter Code which we introduce in the following section.

3 Mathematical Structure and Methods

The methods in this paper all use high-dimensional vectors to represent concepts. There are many ways of generating such representations. Ours is based upon the RI paradigm using terminology as described in [5], in which *semantic vectors* are built as superpositions of randomly generated *elemental vectors*, derived by training over a corpus of documents. Throughout this paper we will write $E(\mathrm{X})$ and $S(\mathrm{X})$ for the elemental and semantic vectors associated with the concept X. In addition to concept vectors, we introduce vectors for relations. For example, $E(\mathrm{R})$ would denote the elemental vector for the relation R. Many relationships are directional, and we will use $\mathrm{R}_{\mathrm{inv}}$ to denote the inverse of R, so that A R B and B $\mathrm{R}_{\mathrm{inv}}$ A carry the same external meaning (though they may in some cases be represented by different vectors).

Kanerva's Binary Spatter Code [16] provides the means to encode typed relations into a high-dimensional binary vector space. The Spatter Code is one of a group of representational approaches collectively known as Vector Symbolic Architectures [17] (VSAs), which originated from Smolensky's tensor product based approach [18], and include Holographic Reduced Representations (HRRs) [19] amongst others. VSAs differ from earlier connectionist representations as they allow for the encoding of typed relations and nested compositional structure. Most of the definitions given below work for VSAs in general. However, we make particular use of VSAs with binary-valued vectors and component-wise exclusive or (XOR) as the binding operation: this has the special property of being its own inverse, which the reader should not assume for other implementations.

The primary operations facilitated by VSAs are *binding* and *bundling*. Binding is a multiplication-like operator through which two vectors are combined to form a third vector C that is dissimilar from either of its component vectors A and B. We will use the symbol "$\otimes$" for binding, and the symbol "$\oslash$" for the inverse of binding throughout this paper. Be aware that binding may have different implementations in different models, and is not meant to be identified with the tensor product. It is important that this operator be invertible: if C = A $\otimes$ B, then A $\oslash$ C = A $\oslash$ (A $\otimes$ B) = B. In some models, this recovery may be approximate, but the robust nature of the representation guarantees that A $\oslash$ C is similar enough to B that B can easily be recognized as the best candidate for A $\oslash$ C in the original set of concepts. Thus the invertible nature of the bind operator facilitates the retrieval of information encoded during the binding process. While this operator varies across VSAs, it results in a product that is of the same dimensionality as the component vectors from which it was derived, unlike the tensor product which has the dimensionality of its component vectors squared. When XOR is used, binding commutes: A $\otimes$ B = B $\otimes$ A.

Bundling is an addition-like operator, through which superposition of vectors is achieved. For example, vector addition followed by normalization is commonly employed as a bundling operator. Unlike binding, bundling results in a vector that is maximally similar to its component vectors. We will write the usual "+" for bundling, and the computer science "+=" for "bundle the left hand side with the right hand side and assign the outcome to the symbol on the left hand side." So for example, $S(\mathrm{A})$ += $E(\mathrm{B})$ means "increment the semantic vector for A by the elemental vector for B using the bundling operator." This in particular is a very standard operation in training.

In the case of the spatter code, XOR is used as a binding operator. As it is its own inverse, the binding and decoding processes are identical ($\otimes$=$\oslash$). For bundling, the spatter code employs a majority vote: if the component vectors of the bundle have more ones than zeros in a dimension, this dimension will have a value of one, with ties broken at random (for example, bundling the vectors 011 and 010 may produce either 010 or 011). Once a vector representation for a concept has been built up by binding and/or bundling, it is possible to apply an operator that reverses the binding process to the vector as a whole.

The XOR operator used in the spatter code offers an apparent advantage over the original permutation-based implementation of PSI: both concepts and relations are represented as high-dimensional binary vectors. This suggests relatively simple ways to direct search across predicate paths of interest, such as those that have been shown useful for literature-based discovery [20]. For example, the "ISA-$\mathrm{TREATS}_{\mathrm{inv}}$" path, which may identify conditions treated by the class a drug belongs to, can be specified as "$S(\mathrm{prozac}) \oslash E(\mathrm{ISA}) \otimes E(\mathrm{TREATS}_{\mathrm{inv}})$." To explore the potential advantages of this formulation, we generated a binary implementation of PSI. This differs from our previous implementation in several ways, summarized in Table 1.

Table 1: Comparison between real vector and binary vector implementation of PSI

Implementation	**Real/Permutation-based**	**Binary**
Semantic vectors $S(\mathrm{X})$	Real vectors ($d = 500$)	Binary vectors ($d = 16{,}000$)
Elemental vectors $E(\mathrm{X})$	Sparse ternary	Dense binary
Represent predicate R	Assign permutation P_R	Assign elemental vector $E(\mathrm{R})$
Reversed predicates $\mathrm{R}_{\mathrm{inv}}$	Use natural inverse P_R^{-1}	Assign new elemental vector $E(\mathrm{R}_{\mathrm{inv}})$
Encoding / training of relationship X R Y	$S(X) \mathrel{+}= P_R(E(\mathrm{Y}))$ $S(\mathrm{Y}) \mathrel{+}= P_R^{-1}(E(\mathrm{X}))$	$S(\mathrm{X}) \mathrel{+}= E(\mathrm{R}) \otimes E(\mathrm{Y})$ $S(\mathrm{Y}) \mathrel{+}= E(\mathrm{R}_{\mathrm{inv}}) \otimes E(\mathrm{X})$
Superposition	Vector addition	Majority vote

We are now in a position to describe our core algorithm for building the binary PSI space used in our experiments throughout the rest of this paper. The procedure is as follows:

1. **Assign an elemental vector *E*(X) to each concept X** that occurs 100,000 times or less in the database. More frequent concepts are excluded as they tend to be uninformative, approximating use of a stop-word list. Elemental vectors are 16,000-dimensional binary vectors with a 50% chance of a one or zero in each position.
2. **Assign an elemental vector *E*(R) to each predicate type R** excluding negations and the PROCESS_OF predicate,[5] which has shown to be uninformative. In most cases, two vectors are assigned, one for each direction of the predicate R and $\mathrm{R}_{\mathrm{inv}}$, to distinguish between the roles of the concepts involved. For a small number of symmetric predicate types, such as COEXISTS_WITH, only one vector is assigned. Note that this process differs from the original implementation using permutations as operations, since each permutation P has a natural distinct inverse P_{-1}. This is not the case for the current implementation, since XOR is its own inverse. In addition we assign a vector "GA" to represent general association.
3. **Assign a semantic vector to each concept** occurring 100,000 or fewer times. In this implementation, semantic vectors contain 16,000 real-valued variables, initially set to zero. These keep track of votes in each dimension to facilitate bundling.
4. **Statistical weighting** is applied to accentuate the influence of infrequent terms. Inverse document frequency (idf) is calculated for concepts and predicates, and applied during encoding such that general associations are weighted according to the idf of the concept concerned, while specific (typed) relations are weighted accord-

[5] This predicate occurs in predications such as "tuberculosis PROCESS_OF patient" which would create an uninformative link between most human diseases.

ing to the sum of the idfs of the concept and the predicate concerned. Consequently, specific relations are weighted more heavily than general relatons.

5. **Process the predications a concept occurs in**: each time a concept occurs in a predication, add (bundle) to its semantic vector the elemental vector for the other concept in the predication bound with the elemental vector for the predicate concerned. For example, when the concept fluoxetine occurs in the predication "fluoxetine TREATS major depressive disorder (MDD)," we add to $S(\text{fluoxetine})$ the elemental vector for TREATS bound with the elemental vector for MDD. We also encode general association by bundling the elemental vector for MDD bound with the elemental vector for general association (GA), ensuring that two concepts relating to the same third concept will have similar vectors, even if they relate to it in different ways. In symbols, we have that $S(\text{fluoxetine})$ += $E(\text{TREATS})$ $\otimes$ $E(\text{MDD}) + E(\text{GA}) \otimes E(\text{MDD})$.

The PSI space was derived from a set of 22,669,964 predications extracted from citations added to MEDLINE over the past decade by the SemRep natural language processing system [21], which extracts predications from biomedical text using domain knowledge in the Unified Medical Language System [22]. For example, the predication "fluoxetine TREATS MDD" is extracted from "patients who have been successfully treated with fluoxetine for major depression." In a recent evaluation of SemRep, Kilicoglu et al. report .75 precision and .64 recall (.69 f-score) [23].

4 Analogical Retrieval

Now that we have built our PSI space, we can use it to search for relations and analogies of concepts as described in the abstract and introduction. The process for performing this search in predication space is similar to Kanerva's XOR-based analogical mapping [2]. Consider the vectors $S(\text{fluoxetine})$ and $E(\text{MDD})$:

$$\begin{aligned} S(\text{fluoxetine}) &= E(\text{MDD}) \otimes E(\text{TREATS}) + E(\text{MDD}) \otimes E(\text{GA}) \\ S(\text{fluoxetine}) \oslash E(\text{MDD}) &= E(\text{MDD}) \oslash E(\text{MDD}) \otimes E(\text{TREATS}) \\ &\quad + E(\text{MDD}) \oslash E(\text{MDD}) \otimes E(\text{GA}) \\ &= E(\text{TREATS}) + E(\text{GA}) \end{aligned}$$

When encoding many predications, the result will be a noisy version of this vector, which should be approximately equidistant from $E(\text{TREATS})$ and $E(\text{GA})$. Therefore we would anticipate being able to search for the treatment for schizophrenia, for example, by finding the semantic vector that is closest to the vector "$S(\text{fluoxetine}) \oslash E(\text{MDD}) \otimes E(\text{schizophrenia})$." This search approximates the single-relation analogies that occur as questions in standardized tests such as the SAT, and have been the focus of recent evaluations of distributional models that estimate relational similarity (eg. [24]). However, useful predicate paths, such as the ISA-TREATS$_{\text{inv}}$ example, often involve more than one relation. The mathematical properties of the binary PSI space suggest that a similar approach can also be used to search across two relations. Consider the following steps that occur during generation of the binary PSI space:

$$S(\text{amoxicillin}) \mathrel{+}= E(\text{antibiotics}) \otimes E(\text{ISA})$$

$$\begin{aligned} S(\text{streptococcal tonsilitis}) &\mathrel{+}= E(\text{antibiotics}) \otimes E(\mathsf{TREATS}_{\text{inv}}) \\ S(\text{prozac}) &\mathrel{+}= E(\text{fluoxetine}) \otimes E(\mathsf{ISA}) \\ S(\text{MDD}) &\mathrel{+}= E(\text{fluoxetine}) \otimes E(\mathsf{TREATS}_{\text{inv}}) \end{aligned}$$

Assuming for the sake of simplicity that these are the only encoding operations that have taken place, an example cue could be generated as follows:

$$\begin{aligned} & S(\text{amoxicillin}) \oslash S(\text{streptococcal tonsilitis}) \\ & \qquad = E(\mathsf{ISA}) \otimes E(\text{antibiotics}) \oslash E(\text{antibiotics}) \otimes E(\mathsf{TREATS}_{\text{inv}}) \\ & \qquad = E(\mathsf{ISA}) \otimes E(\mathsf{TREATS}_{\text{inv}}) \\ & S(\text{MDD}) \oslash S(\text{amoxicillin}) \oslash S(\text{streptococcal tonsilitis}) \\ & \qquad = E(\text{fluoxetine}) \otimes E(\mathsf{TREATS}_{\text{inv}}) \oslash E(\mathsf{TREATS}_{\text{inv}}) \otimes E(\text{ISA}) \\ & \qquad = E(\text{fluoxetine}) \otimes E(\mathsf{ISA}) \\ & \qquad = S(\text{prozac}) \end{aligned}$$

Table 2 illustrates analogical retrieval with single and dual predicates. For single predicates (top three examples), the cue is constructed by combining $E(\text{schizophrenia})$ with the elemental and semantic vector of a pair of concepts, using XOR. The nearest semantic vector to this composite cue is in all cases related to schizophrenia by the same relation that links the example pair: emd_57445 is an experimental treatment for schizophrenia [25], syngr1 is a gene that has been associated with it [26], and certain mannerisms are relevant to the diagnosis of schizophrenia.

Table 2: Schizophrenia-related searches, single- (top 3) and dual-predicate (bottom 3). MDD=Major Depressive Disorder. Scores indicate $1-$normalized hamming distance.

Example pair	**Nearest predicate**	**Nearest neighboring semantic vector**
$S(\text{fluoxetine}) \oslash E(\text{MDD})$	$E(\mathsf{TREATS})$	0.56 $S(\text{emd\_57445})$
$S(\text{apolipoprotein e gene})$ $\oslash E(\text{alzheimer's disease})$	$E(\mathsf{ASSOCIATED\_WITH})$	0.76 $S(\text{syngr1})$
$S(\text{wheezing}) \oslash E(\text{asthma})$	$E(\mathsf{DIAGNOSES})$	0.63 $S(\text{mannerism})$
$S(\text{prozac}) \oslash S(\text{MDD})$	$E(\text{ISA}) \otimes E(\mathsf{TREATS}_{\text{inv}})$	0.54 $S(\text{mazapertine succinate})$
$S(\text{diabetes mellitus}) \oslash$ $S(\text{blood glucose fluctuation})$	$E(\mathsf{TREATS}_{\text{inv}}) \otimes$ $E(\mathsf{CAUSES}_{\text{inv}})$	0.55 $S(\text{impaired job performance})$
$S(\text{chronic confusion}) \oslash$ $S(\text{alzheimer's disease})$	$E(\text{ISA}) \otimes$ $E(\mathsf{COEXISTS\_WITH})$	0.76 $S(\text{acculturation difficulty})$

In the case of dual predicates (bottom three examples), the cue is constructed by combining the semantic vector for schizophrenia with the semantic vectors for a pair of concepts, using XOR. Depression is treated by antidepressants such as prozac. Similarly, schizophrenia is treated by antipsychotic agents, such as mazapertine succinate. Blood glucose fluctuation is a side effect of diabetic treatment, as impaired work performance is a side effect of drugs treating schizophrenia. Finally, chronic confusion occurs in dementias such as Alzheimer's, as acculturation difficulty occurs in psychotic disorders such as schizophrenia.

4.1 Evaluation

To evaluate the single-predicate approach, we extracted a set of test predications from the database using the following procedure. Firstly, a set of candidate predicates was selected. Only predicates meeting the previously-listed constraints for inclusion in our vector space model that occurred one thousand or more times in the data set were considered, leaving a total of 37 predicate types (such as $\mathrm{DIAGNOSES}$). For each of these predicates, fifty predications were randomly selected taking into account the strength of association between the example pair (e.g. $S(\mathrm{wheezing}) \oslash E(\mathrm{asthma})$) and the predicate (e.g. $E(\mathrm{DIAGNOSES})$) such that ten examples were obtained for each predicate that fell into the following ranges of association strength: 0.5211-0.6, 0.61-0.7, 0.71-0.8, 0.81-0.9, 0.91-1.0. We sampled in this manner in order to test the hypothesis that better examples would have a stronger cue-to-predicate association strength, and excluded any example pairs in which this association was less than 0.5211, a value 5SD above the median similarity between a set of 5000 random vectors. Only predicates in which ten examples in each category could be found were tested, resulting in a test set of 1400 predications, fifty per eligible predicate (n=28). For each predicate, every example was tested against every other example pair (n=49) using three approaches summarized in Table 3. 68,600 searches were conducted with each approach. In each case, the nearest semantic vector (e.g. $S(\mathrm{mannerism})$) to the composite cue vector (e.g. $S(\mathrm{wheezing}) \oslash E(\mathrm{asthma}) \otimes E(\mathrm{schizophrenia})$) was retrieved, and tested for occurrence in a predication with the object of the second pair (e.g. schizophrenia), and the same predicate as the example pair (e.g. $\mathrm{DIAGNOSES}$).

To evaluate the paired-predicate approach, we selected fourteen relationship pairs representing predicate paths of interest, including our recurring ISA-$\mathrm{TREATS}_{\mathrm{inv}}$ example, and pairs such as $\mathrm{INHIBITS}$-$\mathrm{CAUSES}_{\mathrm{inv}}$ that are of interest for literature-based discovery [20]. For each pair, we extracted sixty example concept pairs by first selecting for each subject (e.g. prozac) occurring in a relationship of the first type (e.g. ISA) the bridging term (e.g. fluoxetine) and object (e.g. MDD) of the second relationship (e.g. $\mathrm{TREATS}_{\mathrm{inv}}$) with the strongest cue-to-predicate-pair association (similarity between $S(\mathrm{prozac}) \oslash S(\mathrm{MDD})$ and $E(\mathrm{ISA}) \otimes E(\mathrm{TREATS}_{\mathrm{inv}})$). This constraint ensured that it was possible to obtain an adequate number of examples at each cue-to-predicate-pair threshold level. These strongly associated paths were sampled at random, such that sixty example pairs were drawn for each predicate pair, with twenty of these occurring in each of the threshold levels 0.5211-0.6, 0.61-0.7, 0.71-1.0.

Each elemental predicate vector was bound to every other predicate vector, to generate a set of 5,929 paired predicate vectors, such as $E(\mathrm{TREATS}_{\mathrm{inv}}) \otimes E(\mathrm{ISA})$, to use for the dual-relation equivalent of the 2-STEP procedure. This and other procedures used to generate cues for this experiment are shown in Table 3. The major difference from the single-relation approach is the use of the semantic vector for both subject and object of the example pair to generate the cue. Also, the general association step does not require binding, as we would anticipate the semantic vectors for two objects associated with the same subject being similar once constructed. Each of the example pairs (n=60) for each predicate pair was tested with the object of every other example pair in the set (n=59), for a total of 49,560 searches per method.

Table 3: Approaches to cue vector generation. sub_1, obj_1 = subject and object from example pair. Obj2 = test object. E(pred_nearest) = nearest predicate vector ((1) single-predicate) or bound predicate vectors ((2) dual-predicate) to bound example pair. GA = general association

Method	**Bound cue vector**	**Example**
1-STEP (1)	$S(\mathrm{sub}_1) \oslash E(\mathrm{obj}_1) \otimes E(\mathrm{obj}_2)$	$S(\mathrm{fluoxetine}) \oslash E(\mathrm{MDD})$ $\otimes E(\mathrm{schizophrenia})$
2-STEP (1)	$E(\mathrm{pred\_nearest}) \otimes E(\mathrm{obj}_2)$	$E(\mathrm{schizophrenia}) \otimes E(\mathrm{TREATS})$
GA (1)	$E(\mathrm{GA}) \otimes E(\mathrm{obj}_2)$	$E(\mathrm{GA}) \otimes E(\mathrm{schizophrenia})$
1-STEP (2)	$S(\mathrm{sub}_1) \oslash S(\mathrm{obj}_1) \oslash S(\mathrm{obj}_2)$	$S(\mathrm{prozac}) \oslash S(\mathrm{MDD}) \oslash S(\mathrm{schizophrenia})$
2-STEP (2)	$E(\mathrm{pred\_nearest}) \oslash S(\mathrm{obj}_2)$	$E(\mathrm{ISA}) \otimes E(\mathrm{TREATS}_{\mathrm{inv}})$ $\oslash S(\mathrm{schizophrenia})$
GA (2)	$S(\mathrm{obj}_2)$	$S(\mathrm{schizophrenia})$

Approaches to cue generation are summarized in Table 3. The generated cues are intended to be similar to the vector representation of the concept (or concepts) providing a solution to an analogical problem of the form **sub**$_1$ is to **obj**$_1$ as **what** is to **obj**$_2$? 1-STEP cue generation binds the example pair to the target object directly. The 2-STEP approach first finds the nearest predicate vector (single predicates) or bound predicate vectors (dual predicates) to the example pair, and then binds this to the target object. The store of predicate vectors here acts as a "clean-up memory" (Plate 1994 [19], pg 101), removing noise from the approximate representation of the predicate (or pair of predicates) retrieved from the example pair. Finally, as a control, we retrieve the concept that our model associates most strongly with the object when the relation type is not considered (General Association, GA). As an additional control, we repeated both experiments while searching the space of elemental vectors using the elemental vector for the test object, to provide a random baseline. As this failed to produce any correct mappings in the vast majority of cases, the results are not shown.

4.2 Results

The results of the single predicate experiment are shown in Fig. 1 (left). The y-axis shows the mean number of test cases in which the retrieved concept occurred in a predication with the test target in which the predicate matched that linking the example pair. Both the 1-STEP and 2-STEP approaches are sensitive to the strength of association between the example pair and the predicate that links them. As might be expected, an intermediate step utilizing clean-up memory improves performance in the 2-STEP approach, particularly as the cue-to-predicate association drops. These results show that an example concept pair can be used to prime search to retrieve concepts that are related to a cue concept in a particular way, with (2-STEP) or without (1-STEP) retrieving a representation of the relationship concerned. This approach is particularly effective with example pairs that have a strong association to the representation of the predicate of interest. The GA approach retrieves a correct mapping less frequently, and is not sensitive to cue-to-predicate association.

Fig. 1 (right) shows the results of the dual-predicate experiment, which are similar to those for single-relation searches: at stronger cue-to-predicate associations, correct

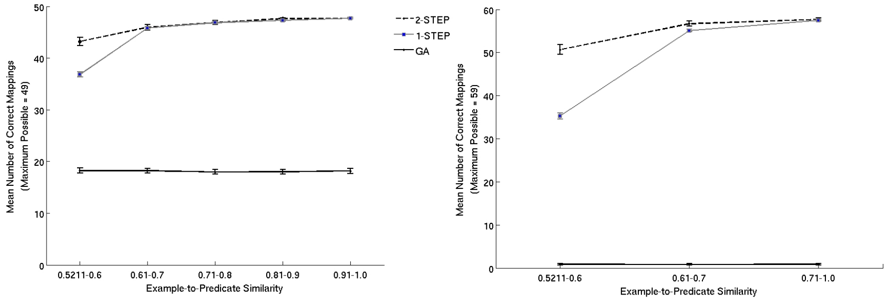

Fig. 1: Analogical retrieval: single (left) and dual (right) predicates. Error bars = standard error.

mappings are found in most cases, whereas with cue-to-predicate associations closer to those anticipated between randomly generated vectors, performance falls. This drop in performance is mitigated to some extent by the use of the 2-STEP approach, in which clean-up memory is used to obtain the original vector representation of the paired relationship concerned. The GA approach is less effective here. While these results do indicate search-by-example is effective in certain cases, the constraint that cue-to-predicate strength should fall in the upper strata limits this approach to a small set of example cues. For example, in the case of the $\mathrm{ISA\text{-}TREATS_{inv}}$ predicate pair, the distribution of cue-to-predicate associations in the set (n=114,658) from which our example cues were sampled (which itself included only the best example for each subject) skews leftward, with a median association strength of 0.522. A similar distribution was observed for single-predicate cues. It is possible to compensate for this using the 2-STEP approach, but this is not ideal for paired relations: with r relations the 2-STEP approach requires searching through r^2 possible predicate pairs. However, as each weak example should have some association with the desired path, we would anticipate the superposition of several weak examples generating a vector with a stronger cue-to-predication-path strength than any of its components. To evaluate this hypothesis, we generated a second set of example pairs for the $\mathrm{ISA\text{-}TREATS_{inv}}$ predicate path. These examples were drawn from the aforementioned set, with the inclusion criterion that their cue-to-predicate association must fall in the weakest category (0.5211 - 0.6). For each example, we measured the cue-predicate association of the example pair $(S(\mathrm{sub_1}) \oslash S(\mathrm{obj_1}))$. As we added new examples, we also measured the association strength between the superposition of all examples up to this point $(S(\mathrm{sub_1}) \oslash S(\mathrm{obj_1}) + \ldots + S(\mathrm{sub_n}) \oslash S(\mathrm{obj_n}))$ and the desired predicate $(E(\mathrm{ISA}) \otimes E(\mathrm{TREATS_{inv}}))$.

The results of this experiment are shown in Fig 2 (left), which shows a rapid rise in cue-to-predicate strength (solid line) as weak examples are added to the superposition. The strength of this association quickly exceeds the cumulative mean (dashed line) association strength of all of the examples added up to that point (individual dots). As shown in Fig. 2 (right), this effect is also observed with respect to performance on the $\mathrm{ISA\text{-}TREATS_{inv}}$ test examples (n=60). This is a particularly important result from the "generalized quantum" point of view. We have used repeated binding and bundling to create a superposition of compound systems that has not been (and probably cannot be) represented as a product of two individual simple systems. In the quantum literature, this phenomenon is known as "entanglement". Thus our experiments demonstrate that several weak example relationships can be superimposed to obtain an entangled representation of the typed relation which is a much more accurate guide for inferring new examples.

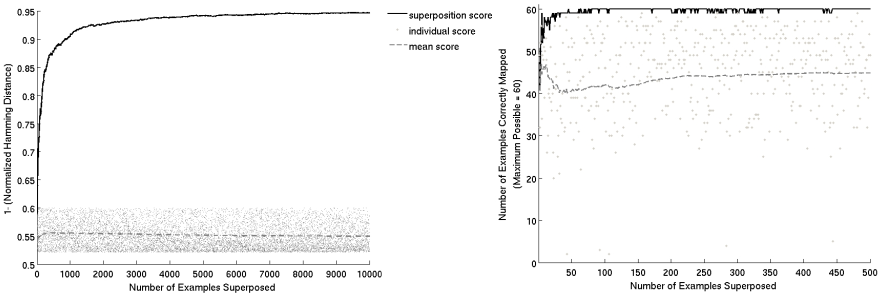

Fig. 2: Superposition: cue-predicate association (left), correct mappings (right).

5 Discussion

In this paper, we show that relational similarity emerges as a natural consequence of the PSI approach. This similarity is sufficient to solve proportional analogy problems stretching across one and two relations, given either a strong example with well-preserved similarity to the relation(s) of interest, or a set of weaker examples. These findings are pertinent to our ongoing research in the area of literature-based discovery and abductive reasoning. Previously, we have discussed various forms of abductive reasoning and constraints operative in such reasoning, and proposed that similarity of some kind is often of importance in finding a link between a starting point of an inquiry and fruitful novel connection to the starting point [27]. The associations are usually weak and indirect, but likely critical in making the connection. Analogy is one form of such indirect connection. An analogy and the starting point have relationships in common [28] so presumably finding cases of common relations is at the heart of analogy retrieval. There have been several implementations of vector encoding to accomplish analogical reasoning [29], [30]. These modeling efforts aim to address several aspects of analogical reasoning: retrieving potential analogies, mapping the elements of the potential target analogy to the elements of the starting point, and making inferences about the starting point from the target analogy. Our goals are more modest in some respects and more ambitious in others. We are initially only concerned with retrieving potential analogies, but we aim to do this on a large scale using large numbers of predications that have been automatically extracted from the biomedical literature, while most of the models of analogies have worked with small sets of custom-constructed predications relating to a few stories. Through analogical retrieval, we are able to direct search across predicate paths that have been shown to be useful for literature-based discovery [20], without incurring an exponential increase in the size of the search space when more than one relationship is considered. The facility for search of this nature is an emergent property of the PSI model: candidates for retrieval are identified on the basis of their similarity to a vector representing a novel relation type, composed from elemental relations during the process of model generation. An approximation of this vector is inferred from the superposition of a set of example pairs, providing an efficient and accurate mechanism for directed search.

6 Conclusion

In this paper, we show that accurate example-based analogical retrieval across single and dual-predicate paths emerges as a natural consequence of the encoding of typed relations in high-dimensional vector space. Given a suitable example pair, or set of less suitable example pairs, it is possible to retrieve with accuracy concepts that relate to another concept in the same way as the concepts in the example pair relate to one another, even if this relationship involves two relations and a third bridging concept. In the case of dual relations, search is achieved without the need to retrieve either the bridging concept or the relations involved. The size of the search space does not increase when dual-relation paths are sought, providing an efficient means to direct predication-based search toward pathways of interest for literature-based discovery.

7 Acknowledgements

This research was supported in part by the US National Library of Medicine grant (R21LM010826-01). The authors would also like to acknowledge Google, Inc. for their support of author DW's ongoing research on the subject.

References

1. T. Cohen, D. Widdows, R. W. Schvaneveldt, and T. C. Rindflesch.: "Logical leaps and quantum connectives: Forging paths through predication space." in AAAI-Fall 2010 Symposium on Quantum Informatics for Cognitive, Social, and Semantic Processes November, pp. 1113, 2010.
2. P. Kanerva.: "Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors." Cognitive Computation, vol. 1, no. 2, pp. 139-159, 2009.
3. C. J. Van Rijsbergen "The Geometry of Information Retrieval." Cambridge University Press, 2004.
4. L. De Vine and P. Bruza.: "Semantic Oscillations: Encoding Context and Structure in Complex Valued Holographic Vectors." Quantum Informatics for Cognitive, Social, and Semantic Processes (QI 2010), 2010.
5. T. Cohen, R. Schvaneveldt, and D. Widdows.:"Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections." Journal of Biomedical Informatics, vol. 43, no. 2, pp. 240-256, Apr. 2010.
6. T. K. Landauer and S. T. Dumais.: "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." Psychological Review, vol. 104, pp. 211-240, 1997.
7. D. R. Swanson.: "Two Medical Literatures that are Logically but not Bib-liographically Connected." Prog Lipid Res, vol. 21, p. 82, 2007.
8. R. A. DiGiacomo, J. M. Kremer, and D. M. Shah.: "Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study." The American journal of medicine, vol. 86, pp. 158-164, 1989.
9. C. S. Peirce.: "Abduction and Induction." in J. Buchler (Ed.) Philosophical writings of Peirce, New York: Routledge, 1940.

10. D. R. Swanson and N. R. Smalheiser.: "An interactive system for finding complementary literatures: a stimulus to scientific discovery." Artificial Intelligence, vol. 91, pp. 183-203, 1997.
11. M. Weeber, J. A. Kors, and B. Mons.: "Online tools to support literature-based discovery in the life sciences." Briefings in bioinformatics, vol. 6, no. 3, pp. 277-286, 2005.
12. P. Kanerva, J. Kristofersson, and A. Holst.: "Random indexing of text samples for latent semantic analysis." Proceedings of the 22nd Annual Conference of the Cognitive Science Society, vol. 1036, 2000.
13. M. D. Gordon and S. Dumais.: "Using latent semantic indexing for literature based discovery." JASIS, vol. 49, pp. 674-685, 1998.
14. P. Bruza, R. Cole, D. Song, and Z. Bari.: Towards Operational Abduction from a Cognitive Perspective, vol. 14. Oxford Univ Press, 2006.
15. T. Cohen, R. Schvaneveldt, and T. Rindflesch.: "Predication-based Semantic Indexing: Permutations as a Means to Encode Predications in Semantic Space." Proceedings of the AMIA annual symposium, San Francisco., 2009.
16. P. Kanerva.: "Binary spatter-coding of ordered K-tuples." Artificial Neural NetworksICANN 96, pp. 869-873, 1996.
17. R. W. Gayler.: "Vector Symbolic Architectures answer Jackendoff's challenges for cognitive neuroscience." In Peter Slezak (Ed.), ICCS/ASCS International Conference on Cognitive Science, pp. 133-138, 2003.
18. P. Smolensky.: "Tensor product variable binding and the representation of symbolic structures in connectionist systems." Artificial intelligence, vol. 46, no. 1, pp. 159-216, 1990.
19. T. A. Plate.: Holographic Reduced Representation: Distributed Representation for Cognitive Structures. CSLI Publications, 2003.
20. D. Hristovski, C. Friedman, T. C. Rindflesch, and B. Peterlin.: "Exploiting semantic relations for literature-based discovery." AMIA Annual Symposium Proceedings, pp. 349-53, 2006.
21. T. C. Rindflesch and M. Fiszman.: "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text." Journal of Biomedical Informatics, vol. 36, pp. 462-477, 2003.
22. O. Bodenreider.: "The unified medical language system (UMLS): integrating biomedical terminology." Nucleic Acids Research, vol. 32, p. D267, 2004.
23. H. Kilicoglu, M. Fiszman, G. Rosemblat, S. Marimpietri, and T. C. Rindflesch.: "Arguments of nominals in semantic interpretation of biomedical text." in Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, pp. 46-54, 2010.
24. P. D. Turney.: "Measuring semantic similarity by latent relational analysis." . Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05), Edinburgh, Scotland, pp. 1136-1141.
25. M. T. Huber, U. Gotthardt, W. Schreiber, and J. C. Krieg.: "Efficacy and safety of the sigma receptor ligand EMD 57445 (panamesine) in patients with schizophrenia: an open clinical trial." Pharmacopsychiatry, vol. 32, no. 2, pp. 68-72, Mar. 1999.
26. R. Verma, S. Kubendran, S. K. Das, S. Jain, and S. K. Brahmachari.: "SYNGR1 is associated with schizophrenia and bipolar disorder in southern India." Journal of Human Genetics, vol. 50, no. 12, pp. 635-640, 2005.
27. R. Schvaneveldt and T. Cohen.: "Abductive Reasoning and Similarity." in In: Ifenthaler D, Seel NM, editor(s). Computer based diagnostics and systematic analysis of knowledge., Springer, New York, 2010.
28. D. Gentner.: "Structure-mapping: A theoretical framework for analogy." Cognitive Science, vol. 7, pp. 155-170, 1983.
29. T. A. Plate.: "Analogy retrieval and processing with distributed vector representations." Expert systems, vol. 17, no. 1, pp. 29-40, 2000.
30. C. Eliasmith and P. Thagard.: "Integrating structure and meaning: A distributed model of analogical mapping." Cognitive Science, vol. 25, no. 2, pp. 245-286, 2001.