Logical Leaps and Quantum Connectives: Forging Paths through Predication Space

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

[1]Center for Cognitive Informatics and Decision Making, School of Health Information Sciences, University of Texas at Houston
[2]Google, inc.
[3]Department of Applied Psychology, Arizona State University
[4]National Library of Medicine
trevor.cohen@uth.tmc.edu

Abstract

The Predication-based Semantic Indexing (PSI) approach encodes both symbolic and distributional information into a semantic space using a permutation-based variant of Random Indexing. In this paper, we develop and evaluate a computational model of abductive reasoning based on PSI. Using distributional information, we identify pairs of concepts that are likely to be predicated about a common third concept, or middle term. As this occurs without the explicit identification of the middle term concerned, we refer to this process as a "logical leap". Subsequently, we use further operations in the PSI space to retrieve this middle term and identify the predicate types involved. On evaluation using a set of 1000 randomly selected cue concepts, the model is shown to retrieve with accuracy concepts that can be connected to a cue concept by a middle term, as well as the middle term concerned, using nearest-neighbor search in the PSI space. The utility of quantum logical operators as a means to identify alternative paths through this space is also explored.

Introduction

The development of alternative approaches to automated reasoning has been a concern of the Quantum Interactions (QI) community since its inception. One line of inquiry has explored the utility of distributional models of meaning as a means of simulating abduction, the generation of new hypotheses, in a computationally tractable manner (Bruza, Widdows, and Woods, 2006). Another concern has been the combination between symbolic and distributional models, and ways in which mathematical models derived from quantum theory might be applied to this end (Clark and Pulman, 2006). This paper describes recent developments along these lines resulting from our work with Predication-based Semantic Indexing (PSI) (Cohen, Schvaneveldt, and Rindflesch, 2009), a novel distributional model that encodes predications, or object-relation-object triplets into a vector space using a variant of the Random Indexing model (Kanerva, Kristofersson, and Holst, 2000). These predications are extracted from citations added to MEDLINE, the most comprehensive database of biomedical literature, over the past decade using the SemRep system (Rindflesch and Fiszman, 2003). We proceed by presenting the methodological roots and implementation of the PSI model, and follow with an discussion of the ways in which abduction can be simulated in the PSI space. Finally, we explore the use of quantum-inspired approaches to concept combination to constrain the process of abduction, with the aim to identify associations between concepts that are of interest for the purpose of biomedical knowledge discovery.

Background

Abduction, Similarity and Scientific Discovery

Abductive reasoning, as defined by the philosopher and logician, C. S. Peirce (1839-1914) is concerned with the generation of new hypotheses given a set of observations. Inductive and deductive reasoning can be applied to confirming and disproving hypotheses, but abductive reasoning is concerned with the discovery of hypotheses as candidates for further testing. Abductive reasoning does not necessarily produce a correct hypothesis, but effective abductive reasoning should lead to plausible hypotheses worthy of further examination and testing. Several factors can be seen to be at work in abductive reasoning (Schvaneveldt and Cohen, 2010). Among these is establishing new connections between concepts. For example, consider information scientist Don Swanson's seminal discovery of a therapeutically useful connection between *Raynaud's disease* and *fish oil* (Swanson, 1986). These concepts had not occurred together in the literature, but were connected to one another by Swanson by identifying potential bridging concepts that did occur with Raynaud's disease (such as *blood viscosity*). Concepts occurring with such bridging concepts were considered as candidates for literature-based discovery. Bruza and his

colleagues note that Swanson's discovery is an example of abductive discovery, and argue that, given the constraints of the human cognitive system, deductive logic does not present a plausible model for reasoning of this nature (Bruza et al., 2006). Rather, associations between terms derived by a distributional model of meaning, in their case Hyperspace Analog to Language (Burgess et al., 1998), are presented as an alternative, a line of investigation we have also pursued in our recent work on literature-based discovery (Cohen, Schvaneveldt, and Widdows, 2009) .

Specifically, we have been concerned with the ability of distributional models to generate *indirect inferences*, meaningful estimates of the similarity between terms that do not co-occur with one another in any document in the database. Such similarities arise because concepts may co-occur with other terms even though they never co-occur with one another. In the context of Swanson's discovery, this would involve identifying a meaningful association between *Raynaud* and *fish oil*. This association would be drawn without the explicit identification of a bridging term. Having identified these associations, it would then be possible to employ some more cognitively and computationally demanding mechanism such as symbolic logic to further investigate the nature of the relationship between these terms. As proposed by Bruza and his colleagues, these associations serve as "primordial stimuli for practical inferences drawn at the symbolic level of cognition" (Bruza, Widdows, and Woods, 2006). The idea that some economical mechanism such as association might be useful in the identification of fruitful hypotheses for further exploration is appealing for both theoretical and practical reasons, the latter on account of the explosion in computational complexity that occurs when considering all possible relations of each potential bridging term in the context of scientific discovery. In addition, there is empirical evidence that associations drawn subconsciously can precede the solution of a problem (Durso, Rea, and Dayton, 1994). In the remainder of this paper, we will discuss the ways in which similarity/association captured by a distributional model of meaning, can support both the identification and validation of hypotheses drawn from the biomedical literature. We begin by presenting some recent technical developments in the field of distributional semantics, to lay the foundation for a discussion of Predication-based Semantic Indexing (PSI) (Cohen et al. 2009), a novel distributional model we have developed in order to simulate aspects of abductive reasoning.

Permutation-based Semantic Indexing

In a previous submission to QI (Widdows and Cohen, 2009), we discussed a recently emerged variant of the RI model developed by Sahlgren and his colleagues (Sahlgren, Holst, and Kanerva, 2008). Based on Pentti Kanerva's work on sparse high-dimensional representations (Kanerva, 2009), this model utilizes a permutation operator that shifts the elements of an elemental vector in order to encode the positional relationship between two terms in a sliding window. In sliding-window based variants of RI, each term is assigned both a sparse *elemental* vector, and a *semantic* vector of a pre-assigned dimensionality several orders of magnitude less than the number of terms in the model (usually on the order of 1000). Elemental vectors consist of mostly zero values, however a small number of these (usually on the order of 10) are randomly assigned as either +1 or -1, to generate a set of vectors with a high probability of being close-to-orthogonal to one another on account of their sparseness. For each term in the model, the elemental vector for every co-occurring term within a sliding window moved through the text is added to the term's semantic vector. The permutation-based model extends this sliding window approach, using shifting of elements in the elemental vector to encode the relative position of terms. Consider the following approximations of elemental vectors:

v1: [-1, 0, 0 , 0 , 1, 0, 0, 0, 1, 0]
v2: [0, -1, 0 , 0 , 0, 1, 0, 0, 0, 1]

Vector v2 has been generated from vector v1 by shifting all of the elements of this vector one position to the right. Of note, these two vectors are orthogonal to one another, and with high-dimensional vectors it is highly probable that a vector permuted in this manner will be orthogonal, or close-to-orthogonal, to the vector from which it is derived. It is also possible to reverse this transformation by shifting the elements one position to the left to regenerate v1. These properties are harnessed by Sahlgren and his colleagues to encode the relative position of terms to one another, providing a computationally convenient alternative to Jones and Mewhort's Beagle model (Jones and Mewhort, 2007), which uses Plate's Holographic Reduced Representation (Plate, 2003) to achieve similar ends. Both of these methodological approaches allow for order-based retrieval. In the case of permutation-based encoding, it is possible, by reversing the permutation used to encode position, to extract from the resulting vector space a term that occurs frequently in a particular position with respect to another term. For example, in a permutation-based space derived from the Touchstone Applied Sciences corpus, the vector derived by shifting the elements of the elemental vector for the term "president" one position to the left produces a sparse vector that is strongly associated with the semantic vectors[1] for the terms "eisenhower", " nixon", "reagan" and "kennedy".

Predication-based Semantic Indexing (PSI)

While the incorporation of additional information related to word order facilitates new types of queries, and has been shown to improve performance in certain evaluations (Sahlgren et al., 2008), the associations derived between terms are general in nature. However, it has been argued that the fundamental unit of meaning in text

[1] It is also possible to use a permuted semantic vector as a cue and search elemental vectors. The differences between these approaches are the subject of ongoing research.

comprehension is not an individual term, but an object-relation-object triplet, or proposition. This unit of meaning is also termed a predication in logic, and is considered to be the atomic unit of meaning in memory in cognitive theories of text comprehension (Kintsch, 1998). While not primarily motivated by cognitive research, the desire to obtain a more constrained measure of semantic relatedness than that provided by cooccurrence-based distributional models has led to the development of wordspace models derived from grammatical relations produced by a parser (Pado and Lapata, 2007). However, these models do not encode the type of relationship that exists between terms, which is desirable for the purpose of mediating scientific discovery as it provides a way of constraining search and simulating cognitive processes involving specific relations.

In our recent work (Cohen, Schvaneveldt and Rindflesch, 2009) we adapt the permutation-based approach developed by Sahlgren *et al* to encode object-relation-object triplets, or predications, into a reduced-dimensional vector space. These triplets are derived from all of the titles and abstracts added to MEDLINE, the largest existing repository of biomedical citation data, over the past decade by the SemRep system (see below). To achieve this end, we assign a sparse elemental vector and a semantic vector to each unique concept extracted by SemRep, and a sequential number to a set of predicate types SemRep recognizes. For example, the predicates "TREATS", "CAUSES" and "ISA" are assigned the numbers 38, 7, and 22 respectively. Rather than use positional shifting to encode the relative position of terms, we use positional shifts to encode the type of predicate that links two concepts. Consequently each time the predication "sherry ISA wine" occurs in the set of predications extracted by SemRep, we shift the elemental vector for the concept "wine" 22 positions to the right, to signify an ISA relationship. We then add this permuted elemental vector to the semantic vector for "sherry". Conversely, we shift the elemental vector for "sherry" 22 positions to the left, and add this permuted elemental vector to the semantic vector for "wine". Encoding predicate type in this manner facilitates a form of predication-based retrieval that is analogous to the order-based retrieval employed by Sahlrgren and his colleagues. For example, permuting the elemental vector for "wine" 22 positions to the right produces a sparse vector with the nearest neighboring semantic vectors and association strengths in Table 1.

Table 1. Results of the predication-based queries "? ISA wine" (left) and "? ISA food" (right).

? ISA wine		? ISA food	
martini	0.73	pastry	0.72
sherry	0.72	rusk	0.72
dry sherry	0.72	dates - food	0.72
fortified wine	0.67	whole grain barley	0.72
wine cooler	0.52	hominy	0.72

Further details of the implementation of this model, and examples of the sorts of queries it enables can be found in (Cohen, Schvaneveldt and Rindflesch 2009). For the purposes of this paper, we have modified the model in order to facilitate the recognition of terms that are meaningfully connected by a bridging term. In PSI, each unique predicate-concept pair is assigned a unique (permuted) elemental vector. Consequently, the semantic vectors for any two concepts should only be similar to one another if they occur in the same predication type with the same bridging concept (discounting unintended random overlap). This constraint is too tight to support scientific discovery, or model abduction. Consequently, in the current iteration of PSI in addition to adding the predicate-appropriate permutation of an elemental vector to the semantic vector of the other concept in a predication, we also add the *unpermuted* elemental vector for this concept. The procedure to encode the predication "sherry ISA wine" would then be as follows. First, add the elemental vector for wine to the semantic vector for sherry. Next, shift the elemental vector for wine right 22 positions and add this to the semantic vector for sherry. The converse would be performed as described previously, but both the permuted and unpermuted elemental vectors for sherry would be added to the semantic vector for wine. Encoding of predicate-specific and general relatedness in this manner is analogous to the encoding of "order-based" and "content-based" relatedness in approaches that capture the relative position of terms (Sahlgren, Holst and Kanerva 2008).

Semrep

The predications encoded by the PSI model are derived from the biomedical literature by the SemRep system. SemRep is SemRep is a symbolic natural language processing system that identifies semantic predications in biomedical text. For example, SemRep extracts "Acetylcholine STIMULATES Nitric Oxide" from the sentence *In humans, ACh evoked a dose-dependent increase of NO levels in exhaled air*. SemRep is linguistically based and intensively depends on structured biomedical domain knowledge in the Unified Medical Language System (UMLS SPECIALIST Lexicon, Metathesaurus, Semantic Network (Bodenreider 2004)). At the core of SemRep processing is a partial syntactic analysis in which simple noun phrases are enhanced with Metathesaurus concepts. Rules first link syntactic elements (such as verbs and nominalizations) to ontological predicates in the Semantic Network and then find syntactically allowable noun phrases to serve as arguments. A metarule relies on semantic classes associated with Metathesaurus concepts to ensure that constraints enforced by the Semantic Network are satisfied.

SemRep provides underspecified interpretation for a range of syntactic structures rather than detailed representation for a limited number of phenomena. Thirty core predications in clinical medicine, genetic etiology of disease, pharmacogenomics, and molecular biology are retrieved. Quantification, tense and modality, and

predicates taking predicational arguments are not addressed. The application has been used to extract 23,751,028 predication tokens from 6,964,326 MEDLINE citations (with dates between 01/10/1999 and 03/31/2010). Several evaluations of SemRep are reported in the literature. For example, in Ahlers et al. (2004) .73 precision and .55 recall (.63 f-score) resulted from a reference standard of 850 predications in 300 sentences randomly selected from MEDLINE citations. Kilicoglu et al. (2010) report .75 precision and .64 recall (.69 f-score) based on 569 predications annotated in 300 sentences from 239 MEDLINE citations. Consequently, the set of predications extracted by SemRep present a considerable resource for biomedical knowledge discovery.

Abduction in PSI-space

For the reasons described previously, the stepwise traversal of all concepts in predications with each middle term that occurs in a predicate with a cue concept is not plausible as a computational model of abduction. Consequently, we have developed a model in which the search for a middle term is guided by an initial "logical leap" from cue concept to target concept.

Our model of abduction consists of the following three stages:

1. Identification of the nearest neighboring *semantic vector* to the semantic vector of a concept of interest.
2. Identification of a third "middle term" between the cue concept and the nearest neighbor. This is accomplished by taking the normalized vector sum (or vector average) of the *semantic vectors* for these two concepts, and finding the most similar *elemental vector*.
3. Decoding of the predicates that link the three concepts identified. For each pair of concepts, this is accomplished by retrieving the *elemental vector* for one, and the *semantic vector* for the other, and shifting one of these by the number corresponding to each encoded predication, to identify the predicate that fits best.

Such "logical leaps" may correspond to an intuitive sense of association in psychological terms. The underlying mechanism may involve associations arising from related patterns of associated neighbors rather than any direct association. These indirect associations are likely to be weaker than direct associations so detecting and reflecting on them may not occur without some effort directed toward searching for potential hypotheses, solutions, or discoveries. Psychological research has provided evidence that such associations occur in learning and memory experiments (Dougher, et al., 1994, 2007; Sidman, 2000). Once detected, indirect associations could be pursued in a more conscious/symbolic way to identify common neighbors or middle terms on the way to assessing the value of the indirect associations. Our computational methods can be seen as ways to simulate the generation and evaluation of such potential discoveries.

In order to evaluate the extent to which this approach can be used to both identify and characterize the nature of meaningful associations, we select at random 1000 UMLS concepts extracted by SemRep from MEDLINE over the past decade. We include only concepts that occur between 10 and 50,000 times in this dataset, to select for concepts that have sufficient data points to generate meaningful associations and eliminate concepts that carry little information content from the test set. We generate a 500 dimensional PSI space derived from all of the predications extracted by SemRep from citations added to MEDLINE over the past decade (n = 22,669.964), excluding negations (x does_not_treat y). We also exclude any predication involving the predicate "PROCESS_OF", as these are highly prevalent but tend to be uninformative (for example, "tuberculosis PROCESS_OF patients"). For the same reason, we exclude any concepts that occur more than 100,000 times in the database.

We then follow the procedure described previously, taking the nearest neighboring semantic vector of each cue concept, generating the vector average of these two vectors, searching for the nearest elemental vector and using the decoding process to find the predication that best links each pair of concepts (cue and middle term, and target and middle term). We then evaluate these predications against the original database, to determine whether these are accurate. Of the 1000 cue concepts it was possible to evaluate 999, as one concept occurred in predications that were not included in the model (such as PROCESS_OF) only. Of these 999 concepts, a legitimate target concept and middle term were identified for 962 of them, which can be considered as a precision of 0.963 if retrieval of a set of accurate relationships from the database is taken as a gold standard. Accurately retrieved results tended to have a higher cosine association between the middle term and the vector average constructed from the cue concept and its nearest neighboring semantic vector, as illustrated in Figure 1, which gives shows the number of accurate and inaccurate results at different association strengths.

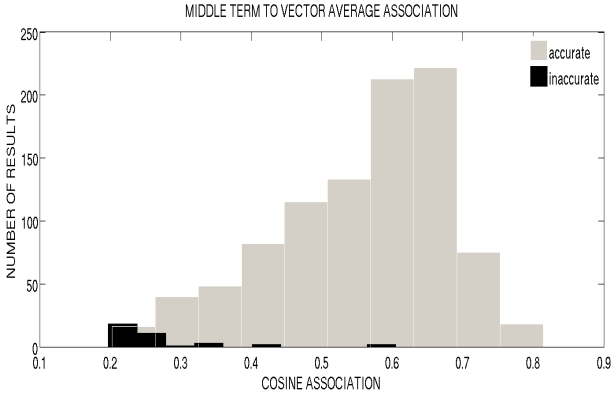

Figure 1: Cosine association and accuracy

Table 2 shows the five most strongly associated middle terms across this test set, together with the predicates

linking them to the cue and target concepts. In the first example, an indirect connection between two molecules has been identified on the basis that they both interact with a third molecule. No predication linking smad proteins to latent tgf-beta binding protein (Ltbp4) occurs in the database. However, a Pubmed search (05/08/2010) for 'smad "latent tgf-beta binding protein"' yields four results. One of these asserts that "a 12-amino-acid deletion in Ltbp4 was associated with increased proteolysis, SMAD signaling, and fibrosis. These data identify Ltbp4 as a target gene to regulate TGF-beta signaling and modify outcomes in muscular dystrophy." (Heydeman et al 2009), which provides support for the inference drawn by the system that these concepts are meaningfully related.

Table 2: "logical leaps". Cue concepts are in bold, and nearest neighbors are underlined.

Cosine -	Cue, neighbor, middle term and predications
0.81 -	**smad_proteins** INTERACTS_WITH transforming_growth_factor_beta ; latent_tgf-beta_binding_protein INTERACTS_WITH transforming_growth_factor_beta
0.81 -	**retinoyl_beta-glucuronide** COEXISTS_WITH - tretinoin ; cyp26a1 INTERACTS_WITH - tretinoin -
0.80 -	**zenapax** true ISA daclizumab ; daclizumab TREATS chronic_orbital_myositis;
0.78 -	**organophosphorus_compounds** INHIBITS acetylcholinesterase ; crotoxyphos INHIBITS acetylcholinesterase
0.78 -	rattus LOCATION_OF **biotinylated_dextran_amine;** rattus LOCATION_OF 1-(methacryloyloxymethyl)propyl hydrogen maleate

In the second example, a connection is drawn between two molecules that do not appear to be discussed together in the literature. Retinoyl beta-glucoronide (RAG) has been shown to have similar biological effects to tretinoin (a form of Vitamin A) but with fewer toxic effects (Barua and Sidell 2004). While it is well established that the protein cyp26A1 acts on tretinoin, a Pubmed search (05/08/2010) for 'cyp26A1 retinoyl beta-glucuronide' produces no results. In the third case, the system has recovered sufficient information to produce a logically consistent answer to the question "does Zenapax treat Chronic Orbital Myositis", despite no predication in the database existing between these two concepts. Zenapax is the trade name of Daclizumab, a therapeutic agent that has been used in the treatment of auto-immune diseases such as orbital myositis. The fourth example has linked both the pesticide "crotoxyphos" and the class of "organophosphorus" compounds it belongs to the their inhibitory effect on acetylcholinesterase, and in the fifth molecules commonly employed in the laboratory are linked through their application to the laboratory rat, "rattus" Norvegicus.

These examples illustrate the ability of vectors encoded using PSI to capture similarity between concepts linked by a middle term without the need to explicitly retrieve this term. However, at times it may be of greater interest to explore some subset of this space, so as to retrieve concepts linked by specific predicate types. One goal of this research is to develop computational tools with which scientists can explore the conceptual territory of their domain of interest. Just as users of a vector-based information retrieval system require methods through which to direct their search for documents, there is a need for the development of methods through which a scientist might further refine the search for new ideas.

Quantum Operators in PSI Space

One potential solution to the problem of constraining search is suggested by the analogy drawn between the many senses of a term that may be captured by a term vector in geometric models of meaning, and the many potential states of a particle that are represented by a state vector in quantum mechanics (Widdows and Peters, 2003). With respect to PSI, the semantic vector representing a concept can be viewed as a mixture of elemental vectors representing each predicate-concept pair and concept it occurs with. This analogy supports the application of the operators of quantum logic, as described by Birkhoff and von Neumann (Birkhoff and Von Neumann, 1936), to semantic vectors, resulting in the definition of semantic space operators effecting quantum logical negation and disjunction in semantic space (Widdows and Peters, 2003).

Negation

Negation in semantic space involves eliminating an undesired sense of a term by subtracting that component of a term vector that is shared with a candidate term representing the undesired sense. For example, the term "pop" can be used to eliminate the musical sense of the term "rock" (Widdows, 2004). This is accomplished by projecting the vector for "rock" onto the vector for "pop" (to identify the shared component), and subtracting this projection from the vector for "rock". The resulting vector will be orthogonal to the vector for "pop", and as such will not be strongly associated with vectors representing music-related concepts that are similar to the vector for "pop", but will retain similarity to terms such as "limestone" that represent the geological sense of "rock".

A similar approach can be applied to the semantic vectors generated using PSI, in order to direct the search for related concepts away from a nearest neighbor that has been identified. As is the case with terms, one would anticipate this approach would eliminate not only the specific concept concerned, but also a set of related concepts. Specifically, we anticipate that this approach

would identify a new path involving a different middle term (or group of terms), without the explicit identification of the middle term to be avoided beforehand.

In order to evaluate the extent to which negation can be used to identify new pathways in PSI space, we take the same set of 1000 randomly selected concepts as cue concepts. For each cue concept, we retrieve the vector for the concept (*cue_concept*), and the vector for the Nearest neighbor previously retrieved (*nn_previous*). We then use negation to extract the component of *cue_concept* that is orthogonal to *nn_previous*, and find the nearest neighboring semantic vector to this combined vector (*nn_current)*. Finally, we take the vector average of *cue_concept* and *nn_current*, render this orthogonal to *nn_previous* using negation, and find the nearest neighboring elemental vector to this combined vector. We then decode the predicates concerned using the permutation operator as described previously.

Table 3: Negation to identify new paths (n=997)

	% new neighbor	% new middle term	% accurate predications
Quantum	100	94.1	92.3
Boolean	n/a	27.7	95.9

The results of this experiment are shown in Table 3. It was possible to obtain results for 997 of the set of 1000. One concept was included for the same reason as previously, and another two were excluded as the negation operator produced a zero vector, as these concepts occurred in predications exclusively with a single predicate-concept pair. As anticipated, in every case negation eliminated the concept represented by *nn_previous*. However, this result could have been obtained using boolean negation, which is the equivalent of simply by selecting the next-nearest neighbor, as we have done for comparison purposes.

Of greater interest is the extent to which the use of quantum negation eliminates the path across a middle term that was used to identify a previous neighbor. This occurred after quantum negation in 94.1% of cases, as oppose to 27.7% in the case of boolean negation. A concern with the use of this method is that the orthogonalization process may introduce further errors as concept vectors are distorted beyond recognition. However, as shown in Table 3, this process led to only slightly more erroneous predications than were obtained with boolean negation. Interestingly, the set of errors produced in the original experiment has very few elements in common with the set produced after quantum negation – erroneous predications were produced for only four of the same cue terms.

Disjunction

We note that it is possible to select for particular predicate types by reversing the permutation operator that corresponds to the predicate of interest. For example, the predication A TREATS B is encoded by shifting the elemental vector for A 38 steps to the right, and adding this to the semantic vector for B. Applying the reverse shift to the semantic vector for B, to produce B^ should produce a vector that retains some remnant of the original elemental vector for A. As both B and B^ should contain remnants of this unpermuted elemental vector, we can isolate concepts that are encoded with this predicate using the following procedure, which we will term *dissection*:

```
for each dimension i:
  if sign B[i] == sign B^[i]:
    BB^[i] = min(absolute_val(B[i]),
                 absolute_val(B^[i]))
  else:
    new_vector[i] = 0
```

Admittedly this is something of a blunt instrument with which to attempt to dissect out remnants of elemental vectors of interest. However, robustness is one advantage of hyper-dimensional representations (Kanerva, 2009), and as illustrated by the results below this method is somewhat successful as a way to isolate desired senses. Once vectors representing the desired sense of a concept have been isolated using this procedure, it is possible to construct a subspace with these vectors as bases. This subspace then represents the set {sense1 OR sense2 OR … sense n} and can be modeled using quantum disjunction, after ensuring the bases of the subspace are orthogonal to one another using the Gram-Schmidt procedure. The association strength between each semantic vector and this subspace can then be measured by projecting a semantic vector into the subspace and measuring the cosine between the original semantic vector and this projection.

To demonstrate the effect of disjunction in PSI space, we construct two subspaces for each concept considered. The first, a biologically oriented subspace, is built from a set of basis vectors that attempt to isolate the following predicates using the procedure described above: AFFECTS, ASSOCIATED_WITH, AUGMENTS, CAUSES, DISRUPTS, INHIBITS, INTERACTS WITH, LOCATION OF, STIMULATES and PART OF. The second, a clinically oriented subspace, attempts to isolate the following predicates with the same procedure: DIAGNOSES, ISA, TREATS, COEXISTS WITH, EVALUATION OF, USES and MANIFESTATION OF.

Table 4 illustrates the effect of quantum disjunction on "logical leaps" in PSI space. A biologically-oriented subspace is generated for each cue concept, and search is conducted by projecting each semantic vector into this subspace, and measuring the cosine between the original semantic vector and its projection. Subsequently, the same procedure is followed, but vector are projected into a clinically-oriented subspace constructed for each cue concept. Any middle term retrieved with an association strength of more than 0.25[1] to the vector average of the cue concept and nearest neighbor is shown in the table.

[1]The strongest between any pair of 1,000 elemental vectors (d=500) (Schvaneveldt, Cohen, and Whitfield, in press)

Table 4: Dissection and Disjunction. PD = Parkinson Disease, AD = Alzheimer's Disease

Cue concept	Nearest neighbor	Middle term(s), **predication(s)**
PD, biological	0.342 multiple system atrophy	0.431 neuro-degenerative disorders (**ISA**) 0.278 alpha-synuclein (**ASSOCIATED_WITH**)
PD, clinical	0.58 huntington disease	0.606 neuro-degenerative disorders (**ISA)**
AD, biological	0.499 iliac artery ectasia	0.659 app (**ASSOCIATED_WITH)**
AD, clinical	0.472 dementia, vascular	0.54 dementia (**ISA)**

The nearest neighbor is influenced by the choice of subspace. While it should be possible to also influence the choice of middle term by projecting the vector average of the cue concept and nearest neighbor into one or the other subspace prior to the middle term search, we have not implemented this here. Consequently, as is the case in the first example, a "clinically" linked middle term may still be retrieved using the "biological" nearest neighbor, but this neighbor is also linked to the cue term by an above-threshold connection through a "biological" middle term. While further evaluation with a larger test set is required, these results suggest the combination between dissection and disjunction might form the basis for a selective search strategy. In both cases, the biologically oriented subspace is close to a neighbor linked through a middle term via a predicate, ASSOCIATED WITH, from the biological set. Likewise, the clinically-oriented subspaces are close to concepts linked appropriately through a middle term via a predicate from the clinical set, ISA. The interpretation of the biological results is interesting also. Multiple System Atrophy is a neurodegenerative disorder that shares several symptoms with PD, and like PD involves the accumulation of abnormal protein aggregates in nerve cells. These are called Lewy Bodies, and Alpha-synuclein is one of their major components (Tong et al., 2010). While the association between AD and the Amyloid Precursor Protein (app) is well established, the predication Iliac Artery Ectasia (abnormal dilatation of a major artery in the pelvis) ASSOCIATED_WITH AD was extracted by SemRep from the following sentence:"Significant CIA ectasia or small aneurysm is often associated with AAA." (Kritpracha et al., 2002). The acronym AAA can refer to both Amyloid of Aging and Alzheimer's and Abdominal Aortic Aneurysm, and this error most likely arose on account of the concept recognition component of SemRep selecting the contextually inappropriate alternative.

Combining Negation and Disjunction

While these experiments do suggest that combining dissection and disjunction may be an effective way to constrain search according to particular predicates, the disjunction operator has the computational disadvantage of requiring each candidate vector in a search to be compared with each component vector of the subspace. This disadvantage can be overcome by the combination of vector negation and disjunction (Widdows and Peters, 2003). When these operators are combined, a cue vector can is made orthogonal to all of the vectors in a disjoined subspace. Consequently, for biologically-oriented searches we now generate a vector for the cue concept that is orthogonal to the clinically oriented subspace, and vice-versa. While not included here, we note that the results obtained with this approach were similar to those shown in Table 4. As search with this approach only requires vector-to-vector comparison, it presents an appealing alternative.

Conclusion

In this paper, we develop and evaluate a model of automated reasoning based on "logical leaps", in which meaningful associations between concepts learned from distributional statistics are used to identify candidates for connection via a third concept, and a symbolic approach is used to identify the nature of the relations involved. On account of its economy, this approach is appealing for theoretical and practical reasons. Furthermore, the vector spaces used for these experiments can be retained in RAM to facilitate rapid, dynamic, interactive exploration of biomedical concepts to support discovery. Vector operators derived from quantum logic show promise as a means to direct such searches away from previously trodden paths, and exploratory work suggests there may be ways to adapt these operators to guide search toward conceptual territory of interest. Of particular interest for future work is the evaluation of the extent to which these operators might be used to model "discovery patterns" (Hristovski, Friedman and Rindflesch 2008), combinations of predications that have been shown useful for literature-based discovery.

References

Ahlers, C. B., Fiszman, M., Demner-Fushman, D., Lang, F-M., & Rindflesch, T. C. (2007). Extracting semantic predications from Medline citations for pharmaco-genomics. *Pacific Symposium on Biocomputing* 12:209-20.

Barua, A. B., and Sidell, N. (2004). Retinoyl {beta} Glucuronide: A Biologically Active InterestingRetinoid. *J. Nutr.*, *134*(1), 286S-289.

Birkhoff, G., and Von Neumann, J. (1936). The Logic of Quantum Mechanics. *The Annals of Mathematics*, Second Series, *37*(4), 823-843.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research* 32 (Database issue): D267-70.

Bruza, P., Cole, R., Song, D., & Bari, Z. (2006). *Towards Operational Abduction from a Cognitive Perspective* (Vol. 14). Oxford Univ Press.

Bruza, P. D., Widdows, D., & Woods, J. (2006). A Quantum Logic of Down Below. In: Engesser, K. Gabbay, D. Lehmann, D. (eds). Handbook of quantum logic and quantum structrures: quantum logic. Elsevier. p. 625-60

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in Context Space: Words, Sentences, Discourse. *Discourse Processes*.

Clark, S., & Pulman, S. (2006). Combining symbolic and distributional models of meaning. In *AAAI Spring Symposium on Quantum Interaction*.

Cohen, T., Schvaneveldt, R., & Rindflesch, T. (2009). Predication-based Semantic Indexing: Permutations as a Means to Encode Predications in Semantic Space. *Proceedings of the AMIA annual symposium, San Francisco.*

Cohen, T., Schvaneveldt, R., & Widdows, D. (2010). Reflective Random Indexing and Indirect Inference: A Scalable Method for Discovery of Implicit Connections. *Journal of Biomedical Informatics*, 43, 2, 240-256.

Dougher, M. J., Auguston, E., Markham, M. R., Greenway,D. E., & Wulfert, E. (1994). The transfer of respondent eliciting and extinction functions through stimulus equivalence classes. *Journal of the Experimental Analysis of Behavior*, 62, 331-351.

Dougher, M. J., Hamilton, D. A., Fink, B. C., & Harrington, J. (2007). Transformation of the discriminative and eliciting functions of generalized relational stimuli. *Journal of the Experimental Analysis of Behavior*, **88**, 179-197.

Durso, F. T., Rea, C. B., & Dayton, T. (1994). Graph-Theoretic Confirmation of Restructuring During Insight. *Psychological Science*, *5*(2), 94-98.

Engelmore, R., and Morgan, A. eds. 1986. *Blackboard Systems.* Reading, Mass.: Addison-Wesley.

Heydemann, A., Ceco, E., Lim, J. E., Hadhazy, M., Ryder, P., Moran, J. L., Beier, D. R., et al. (2009). Latent TGF-beta-binding protein 4 modifies muscular dystrophy in mice. *The Journal of Clinical Investigation*, *119*(12), 3703-3712.

Hristovski, Dimitar; Carol Friedman; and Thomas C. Rindflesch. (2008). Literature-based discovery using natural language processing. Bruza, P and Weeber, M (eds.) *Literature-based Discovery*. Berlin: Springer-Verlag, 153-72.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1-37.

Kanerva, P. (2009). Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation*, *1*(2), 139–159.

Kanerva, P., Kristofersson, J., & Holst, A. (2000). Random indexing of text samples for latent semantic analysis. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, *1036*.

Kilicoglu, H., Fiszman, M., Rosemblat, G., Marimpietri, S., & Rindflesch, T. C. (2010). Arguments of nominals in semantic interpretation of biomedical text. *Proceedings of the BioNLP Workshop*. Association for Computational Linguistics.

Kintsch, W. (1998). *Comprehension : a paradigm for cognition*. Cambridge, ; New York, NY: Cambridge University Press.

Kritpracha, B., Pigott, J. P., Russell, T. E., Corbey, M. J., Whalen, R. C., DiSalle, R. S., Price, C. I., et al. (2002). Bell-bottom aortoiliac endografts: an alternative that preserves pelvic blood flow. *Journal of Vascular Surgery: Official Publication, the Society for Vascular Surgery [and] International Society for Cardiovascular Surgery, North American Chapter*, *35*(5), 874-881.

Pado, S., & Lapata, M. (2007). Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, *33*, 161-199.

Plate, T. A. (2003). *Holographic Reduced Representation: Distributed Representation for Cognitive Structures*. CSLI Publications.

Rindflesch, T. C., & Fiszman, M. (2003). The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics*, *36*, 462-477.

Schvaneveldt, R, Cohen, T and Whitfield, K. Paths to Discovery. *To Appear In: Proceedings of the 36th Carnegie Mellon Symposium on Cognition. June, 2009.*

Sahlgren, M., Holst, A., & Kanerva, P. (2008). Permutations as a Means to Encode Order in Word Space. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society (CogSci'08), July 23-26, Washington D.C., USA.*

Schvaneveldt, Roger, & Cohen, Trevor. (2010). Abductive Reasoning and Similarity. In *In: Ifenthaler D, Seel NM, editor(s). Computer based diagnostics and systematic analysis of knowledge.* Springer, New York.

Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect Biol Med*, *30*(1), 7-18.

Sidman, M. (2000). Equivalence relations and the reinforcement contingency. *Journal of the Experimental Analysis of Behavior*, **74**, 127-146.

Tong, J., Wong, H., Guttman, M., Ang, L. C., Forno, L. S., Shimadzu, M., Rajput, A. H., et al. (2010). Brain alpha-synuclein accumulation in multiple system atrophy, Parkinson's disease and progressive supranuclear palsy: a comparative investigation. *Brain: A Journal of Neurology*, *133*(Pt 1), 172-188. doi:10.1093/brain/awp282

Widdows, D., and Cohen, T. (2009). Semantic Vector Combinations and the Synoptic Gospels. *Proceedings of the Third Quantum Interaction Symposium (March 25-27, 2009 - DFKI, Saarbruecken)*.

Widdows, D., & Peters, S. (2003). Word Vectors and Quantum Logic Experiments with negation and disjunction. *Mathematics of Language, 8, Bloomington, Indiana, June 2003*.

Widdows, D. (2004). *Geometry and Meaning*. Center for the Study of Language and Information/SRI.