Abduction in Semantic Space: Towards a Logic of Discovery

PETER BRUZA, *Distributed systems Technology Centre, University of Queensland, Australia. 4072. Email: bruza@dstc.edu.au*

DAWEI SONG, *Distributed Systems Technology Centre, University of Queensland, Australia. 4072. Email: song@dstc.edu.au*

ROBERT McARTHUR, *Distributed Systems Technology Centre, University of Queensland, Australia. 4072. Email: mcarthur@dstc.edu.au*

Abstract

Diminishing awareness is a consequence of the information explosion: disciplines are becoming increasingly specialized; individuals and groups are becoming ever more insular. This paper considers how awareness can be enhanced via abductive knowledge discovery the goal of which is to produce suggestions which can span disparate islands of knowledge. Knowledge representation is motivated from a cognitive perspective. Words and concepts are represented as vectors in a high dimensional semantic space automatically derived from a text corpus. Information flow computation between vectors is proposed as a means of suggesting potentially interesting implicit associations between concepts. Information flow is applied to computational scientific discovery by attempting to simulate Swanson's Raynaud-fish oil discovery in medical texts. Both automatic and semi-automatic attempts were studied and compared against suggestions computed via semantic association. Even though this work is preliminary and speculative in nature, there is some justification to believe that appropriate suggestions (hypotheses) can be "abduced" from semantic space.

Keywords: Abduction, knowledge discovery, scientific discovery

1 Introduction

In the mid-nineteen eighties, Don Swanson, a librarian, made a chance discovery by connecting two disparate on-line medical literatures, one dealing with Raynaud's disease, the other with fish oil. Patients with Raynaud's disease suffer from intermittent blood flow in the extremities (fingers, toes, ears). At the time, there was neither a general treatment, nor cure, for this disease. Through this chance discovery, Swanson formulated the hypothesis that fish oil may be a cure, which was later verified by clinical trials. Swanson noted, "the two literatures are mutually isolated in that the authors and the readers of one literature are not acquainted with the other, and vice versa." ([23], p184). Were these communities aware of each other, a cure would probably been found much earlier than Swanson's serendipitous discovery. Swanson's observation is an example of a more widely occurring phenomenon: due to the flood of information, disciplines and expertise are becoming increasingly specialised with little awareness of kindred, or potentially allied, specialisations. Automated or semi-automated knowledge discovery systems can counter this growing lack of awareness by

discovering potentially relevant connections between disparate islands of knowledge.

If Swanson's discovery was replicated automatically, what knowledge representation would be appropriate, and how could the hypothesis be generated? According to the philosopher C.S. Peirce, Swanson's explanatory hypothesis is a manifestation of abduction: "It [abduction] is the only logical operation which introduces any new idea; for induction does nothing but determine a value and deduction merely evolves the necessary consequences of a pure hypothesis" ([17], p216). Abduction has recently been considered from a psychologistic perspective which does not permit the reasoning process to be abstracted from the (human) agent performing the reasoning [6]. As abduction is a form of human reasoning, treating it from a psychologistic perspective is particularly apt. The abductive logic system of Gabbay and Woods [6] comprises three components: a logic of discovery, a logic of justification, and a revision component. The logic of discovery uncovers hypotheses, which must then be scrutinized by the logic of justification. Revision of hypotheses occurs when the abduction process goes "bad".

This article will focus on the logic of discovery within the following setting. The raw knowledge will be textual and a psychologistic stance will be adopted with respect to its representation and the discovery of implicit connections in knowledge. These constraints present challenges. Text cannot easily be automatically rendered into a propositional knowledge representation, and as our goal is to build abductive logic systems which can reason over large amounts of text, we don't consider propositional representation of the knowledge to be feasible. In addition, adoption of a psychologistic perspective suggests a cognitively motivated representation of knowledge. Our approach to knowledge representation will be to compute a high dimensional semantic space from the text. Semantic spaces have a demonstrated track record of cognitive compatibility with human information processing, for example, semantic and associative word priming (see later for references). For this reason a semantic space would seem to be a promising basis on which to build a computational system which mimics human reasoning like abduction. Moreover, semantic spaces have been constructed from very large collections of text, for example, a corpus of Usenet news comprising 160 million words [15, 4], so they have a demonstrated track record of knowledge representation in the large.

Finally, there is a need to address C.S. Peirce's comment that abduction produces something "new". Two mechanisms will be presented in this regard; both compute implicit associations between words or concepts. The first mechanism uncovers implicit associations by computing information flow within the high dimensional semantic space. The second technique computes semantic associations. The logic of discovery is evaluated by attempting to replicate Swanson's discovery by automatic means.

2 Semantic knowledge representation via HAL

To bridge the gap between cognitive knowledge representation and actual computational representations, the Hyperspace Analogue to Language (HAL) model is proposed [15, 4]. HAL produces representations of words in a high dimensional space that seem to correlate with the equivalent human representations. For example, "...simulations using HAL accounted for a variety of semantic and associative word priming effects that can be found in the literature...and shed light on the nature of the word

Dimension	*Value*
nifedipine	0.44
scleroderma	0.36
ketanserin	0.22
synthetase	0.22
sclerosis	0.22
thromboxane	0.22
prostaglandin	0.22
dazoxobin	0.21
E1	0.15
calcium	0.15
vasolidation	0.15
platelet	0.15
...	...
platelets	0.07
blood	0.07
viscosity	0.07
vascular	0.07
...	...

TABLE 1. Example HAL representation of the concept "Raynaud"

relations found in human word-association norm data"[15]. Given an n-word vocabulary, the HAL space is an $n \times n$ matrix constructed by moving a window of length l over the corpus by one word increment ignoring punctuation, sentence and paragraph boundaries. All words within the window are considered as co-occurring with each other with strengths inversely proportional to the distance between them. Each row i in the matrix represents accumulated weighted associations of word i with respect to other words which preceded i in a context window. Conversely, column i represents accumulated weighted associations with words that appeared after i in a window.

In the experiments reported later, the row and column in the HAL matrix corresponding to a given word i are added to produce a single vector representation for that word. The vector is then normalized to unit length. The set of such representations is termed a *HAL space*. For example, table 1 shows part of the normalized HAL vector for the word "Raynaud" computed by applying the HAL method to a collection of medical texts originating from the MEDLINE collection (the dimensions are ordered on decreasing strength of association). This example demonstrates how a word is represented as a weighted vector whose dimensions comprise other words. The weights represent the strengths of association between "Raynaud" and other words seen in the context of the sliding window: the higher the weight of a word, the more it has lexically co-occurred with "Raynaud" in the same context(s).

The quality of HAL vectors is influenced by the window size: the longer the window, the higher the chance of representing spurious associations between terms. A window size of eight or ten has been used in various studies [15, 4, 3]. More formally, a concept[1] c is a dimensional representation: $c = \langle w_{cp_1}, \ldots, w_{cp_n} \rangle$ where $p_1, \ldots, p_n$ are dimensions of c, n is the dimensionality of the HAL space, and w_{cp_i} denotes the weight of dimension $p_i, 1 \leq i \leq n$ within the representation c. A dimension p_i of c is termed

[1] The term "concept" is used loosely to emphasize that a HAL space is a primitive realisation of a conceptual space

a termed a *quality property* if and only if its weight w_{cp_i} is greater than a no-zero threshold value δ. $\mathcal{QP}_\delta(c)$ denotes the set of quality properties of concept c. $\mathcal{QP}_\mu(c)$ signifies the set of quality properties above mean value in c, and $\mathcal{QP}(c)$ is short for $\mathcal{QP}_0(c)$.

HAL is an exemplar of a growing ensemble of computational models emerging from cognitive science, which are generally referred to as *semantic spaces* [15, 4, 13, 14, 10, 11, 16, 12, 18]. Even though there is ongoing debate about specific details of the respective models, they all feature a remarkable level of compatibility with a variety of human information processing tasks such as word association. Semantic spaces provide a geometric, rather than propositional, representation of knowledge. They can be considered to be approximations of conceptual space proposed by Gärdenfors [7].

Within a conceptual space, knowledge has a dimensional structure. For example, the property colour can be represented in terms of three dimensions: hue, chromaticity, and brightness. Gärdenfors [7] argues that a property is represented as a convex region in a geometric space. In terms of the example, the property "red" is a convex region within the tri-dimensional space made up of hue, chromaticity and brightness. The property "blue" would occupy a different region of this space. A domain is a set of integral dimensions in the sense that a value in one dimension(s) determines or affects the value in another dimension(s). For example, the three dimensions defining the colour space are integral since the brightness of a colour will affect both its saturation (chromaticity) and hue. Gärdenfors extends the notion of properties into concepts, which are based on domains. The concept "apple" may have domains taste, shape, colour etc. Context is modelled as a weighting function on the domains, for example, when eating an apple, the taste domain will be prominent, but when playing with it, the shape domain will be heavily weighted (i.e., it's roundness). Observe the distinction between representations at the symbolic and conceptual levels. At the symbolic level "apple" can be represented as the atomic proposition $apple(x)$, however, within a conceptual space (conceptual level), it has a representation involving multiple inter-related dimensions and domains. Colloquially speaking, the token "apple" (symbolic level) is the tip of an iceberg with a rich underlying representation at the conceptual level. Gärdenfors points out that the symbolic and conceptual representations of information are not in conflict with each other, but are to be seen as "different perspectives on how information is described".

Barwise and Seligman [1] also propose a geometric foundation to their account of inferential information content via the use of real-valued state spaces. In a state space, the colour "red" would be represented as a point in a tri-dimensional real-valued space. For example, brightness can be modelled as a real-value between white (0) and black (1). Integral dimensions are modelled by so called observation functions defining how the value(s) in dimension(s) determine the value in another dimension. Observe that this is a similar proposal, albeit more primitive, to that of Gärdenfors as the representations correspond to points rather than regions in the space.

A HAL representation is an approximation of a Barwise and Seligman state space whereby the dimensions are words. A word, or combination of words, like a noun compound are represented as a point in the space. This point represents the "state" in the context of the associated text collection from which the HAL space is derived. If the collection changes, the state of the word may also change. HAL, however, does make provision for integral dimensions. (We will return to this apparent deficiency

later). Despite a HAL space being an approximation of a conceptual space, it nevertheless has a demonstrated track record of cognitive compatibility [15, 4]. Moreover, HAL spaces can be constructed by a simple algorithm. In short, HAL, and more generally semantic spaces, are a promising, pragmatic means for knowledge representation based on text. Moreover, due to their cognitive credentials, semantic spaces would seem to be a fitting foundation for realising computational variants of human reasoning, like abduction.

3 Practical reasoning with text: information inference and abduction

Gabbay and Woods [5] have recently proffered a notion of cognitive economy founded on compensation strategies employed by a practical agent to alleviate the consequences of its limited resources. Practical reasoning is reasoning performed by practical agents, such as individuals (e.g., Swanson). Practical agency is conceived of in "terms of the degree of access to key cognitive resources such as *information*, *time*, and *computational capacity*". We briefly recount aspects these compensation strategies and then dovetail them into a discussion specifically centered on the cognitive economics surrounding textual information processing.

One compensation strategy employed by agents is to divide reality into natural kinds. Paired with natural kinds is hasty generalization which is based on small sample sizes of natural kinds, and are defeasible in the light of new experience. Hasty generalizations are a compensation strategy for the scarcity of time and information, but are also fallible. With regard to this point, Gabbay and Woods ([5], p18) observe, "If generic inferences from natural kind samples are not quite right, at least they don't kill us. They don't even keep us from prospering".

Partitioning the information space into categories, for example, via taxonomies, can also be considered a compensation strategy: random search taxes heavily the agent's limited resources. The categories which partition the information space are akin to the natural kinds mentioned earlier. In addition, an agent will make hasty judgments in regard to what the information is, or is not, about. Such judgments are somewhat akin to the hasty generalizations mentioned in relation to natural kinds. By way of illustration, consider the short text "Penguin Crossing Bed and Breakfast". Most of us would conclude quickly that this text is not about birds. In regard to the following text, "Linux Online: Why Linus chose a Penguin", human agents with the requisite background knowledge can infer that "Linus" refers to "Linus Torvalds", the inventor of Linux, and the penguin mentioned here has to do with the Linux logo. In relation to "Penguin Books UK", the judgment would likely be that this text is about a publisher. Finally, "Surfing the Himalayas" may lead some agents to conclude the text refers to snowboarding. In short, human beings can generally make robust judgments about what information fragments are, or are not about, even when the fragments are brief, or incomplete. The process of making such judgments will be referred to as *information inference*. We feel that such inference has a decidedly abductive character.

3.1 Inference and Information

The above examples attempt to illustrate that information inference is a very real phenomenon. It is a commonly occurring, though often unnoticed, part of our daily information processing tasks, for example, the perusal of email subject headings, or document summaries retrieved by a search engine. We make hasty information inferences within such tasks because the full processing of the information taxes our limited time and cognitive resources. Barwise and Seligman [1] have formalized the interplay between inference and information in the following way:

> **Inferential Information Content:** To a person with prior knowledge k, *r being F* carries the information that *s is G*, if the person could legitimately infer that *s is G* from *r being F* together with k (but could not from k alone)

Barwise and Seligman illustrate inferential information content with examples of physical situations. For example, *switch being on* carries the information that *the light bulb is lit*, given a suitable background knowledge k. It is instructive to see how the above definition functions with respect to the information inferences drawn from the example text fragments given previously.

Information inferences are often made on the basis of certain words appearing in the context of other words. By way of illustration, consider once again the text fragment, "Linux Online: Why Linus chose a Penguin" and assume the background knowledge k includes "Linus Torvalds invented Linux. The Linux logo is a penguin, etc.". The above definition can be applied as follows: *"Linus" being (together with) "Linux" (in the same context)* carries the information that *"Linus" is "Linus Torvalds"* . Analogously, *"Penguin" being (together with) "Books"* carries the information that *"Penguin" is (a) "publisher"*. On the basis of these examples, Barwise and Seligman's definition of inferential information content would seem to be a promising and apt foundation on which to build an account of information inference. It is therefore important to consider this definition more closely. The striking aspect of this definition is its psychologistic stance, meaning that the inference process is not considered independent of the human agent drawing the information inferences . Barwise and Seligman state in this respect, "... by relativizing information flow to human inference, this definition makes room for different standards in what sorts of inferences the person is able and willing to make" ([1], p23). This position poses an immediate and onerous challenge due to the inherent flexibility that must be catered for. In our case, we restrict our attention to those sorts of inferences which can be drawn on the basis of words seen in the context of other words under the proviso that such inferences correlate with corresponding human information inferences (thereby being faithful to our psychologistic stance).

A second remarkable feature of the above definition is the role played by background knowledge k. "The background theory k of the agent comes much more actively into this account. It is not just there as a parameter for weeding out possibilities, it becomes a first-class participant in the inference process" ([1], p23). Again this poses a challenge to implementing an information inference system. How will k be acquired, used appropriately, and kept up-to-date? In this article, the semantic space constructed by HAL will fill the role of background knowledge k. Recall that background knowledge realized via HAL is not static. As the underlying corpus of text changes,

so to do the semantic representations in the space[2].

3.2 Discovering hypotheses by computing information flow

The appeal of Gärdenfors' cognitive model for our research is that it allows information inference to be considered not only at the symbolic level, but also at the conceptual (geometric) level. In particular, Gärdenfors ([7], p48) conjectures that most of scientific theorizing (i.e., abduction) takes place within the conceptual level. Inference at the symbolic level is typically a linear, deductive process. Within a conceptual space, inference takes on a decidedly associational character because associations are often based on similarity (e.g., semantic similarity), and notions of similarity are naturally expressed within a dimensional space. It may well be that because such associations are formed below the symbolic level of cognition, significant cognitive economy results. This is not only interesting from a cognitive point of view, but also opens the door to providing a computationally tractable logic of discovery. Our assumption is that hypotheses can be generated by computing implicit associations within a semantic space. Here "implicit" echoes C.S Peirce's view that abduction produces "new ideas". If an association is explicit it is, in our view, less likely to be new enough to be interesting.

In this article, "new ideas" will be realized via computations of information flow in semantic space. The motivating theory for information flow stems from Barwise and Seligman's account of inferential information content within a state space [1]. For example, $\mathsf{penguin}, \mathsf{books} \vdash \mathsf{publisher}$ denotes that the concept "publisher" is carried *informationally* by the combination of the concepts "penguin" and "books" [19, 20]. Said otherwise, "publisher" *flows* informationally from "penguin" and "books". Such information flows are determined by an underlying information state space constructed by HAL. In other words, the HAL representation vector for "penguin" is interpreted to be the "state" of that word, modulo the underlying text corpus from which the semantic space was constructed. If this corpus changes, so to do the "states" of the words. The degree of information flow is directly related to the degree of inclusion between the respective information states. Total inclusion leads to maximum information flow. Inclusion is a relation $\succcurlyeq$ over HAL representations [20]. HAL-based information flow is defined as follows:

DEFINITION 3.1

$$i_1, \ldots, i_k \vdash j \quad \text{iff} \quad \mathsf{degree}(\bigoplus_{i=1}^{k} c_i \succcurlyeq c_j) > \lambda$$

where c_i denotes the HAL representation of concept i, and $\bigoplus_{i=1}^{k} c_i$ refers to the combination of the HAL representations $c_1, \ldots, c_k$ into a single representation in the HAL space. This single vector represents the combined concept comprising words $i_1, \ldots, i_k$. Combined concepts often manifest as noun compounds, for example, "Penguin books" or "fish oil". In the experiments reported below, information flow will be computed from a single term ($k = 1$). Therefore, the concept combination will

[2] HAL need simply be re-run on the new corpus yielding the updated semantic space

not play a role as, trivially, $(\bigoplus_{i=1}^{1} c_i = c_1)$. See Gärdenfors [7] for a detailed discussion on concept combination from a cognitive perspective. Details of a concept combination heuristic within HAL spaces can be found be in [20]. The value λ is a empirically determined threshold which governs the minimum amount of information flow necessary to support the symbolic inference relation.

The degree of inclusion is a normalized score computed in terms of the ratio of intersecting quality properties of c_i and c_j to the number of quality properties in the source concept representation c_i:

$$\mathsf{degree}(c_i \succcurlyeq c_j) = \frac{\sum_{p_x \in (\mathcal{QP}_\delta(c_i) \cap \mathcal{QP}(c_j))} w_{c_i p_x}}{\sum_{p_x \in \mathcal{QP}_\delta(c_i)} w_{c_i p_x}} \tag{3.1}$$

The underlying idea of this definition is to make sure that a majority of quality properties of representation c_i appear in representation c_j. The quality properties in the source concept are defined by the threshold δ. For example, in experiments to automatically infer query expansion terms via information flow computations, setting δ to the average dimension weight in the source concept representation produced best results [3]. Note that information flow produces truly inferential character, i.e., dimension j may have a weight of zero in the representation c_i meaning that terms i and j never explicitly co-occurred in the same window (context) during the construction of the semantic space, however, there may be positive information flow from c_i to c_j. This phenomenon is termed *implicit information inference.*

The HAL-based information flow model has been successfully applied to automatic query expansion for document retrieval with encouraging results [3]. Aspects of this work are briefly recounted as it is relevant to considering abduction within semantic space. The key to effective query expansion is the ability to infer expansion terms relevant to the topic of the query. If the inference mechanism is "unsound", then terms extraneous to the query topic will be introduced causing irrelevant documents to be retrieved. This causes a loss of precision[3]. By way of illustration, consider the query space program (TREC[4] topic 011). Information flow based query expansion inferred[5] terms such as "NASA", "shuttle", "rocket", "satellite", "Soviet" etc. The analysis reported in [3] states that query expansion terms derived via implicit information inference contribute most to increased precision.

Query expansion can also be viewed as an abductive process. The task is to suggest (i.e., "abduce") terms relevant to the topic of the query. Terms which exhibit high degrees of information flow from the given query can be considered, collectively, as furnishing explanatory hypotheses with regard to the query at hand, modulo the underlying semantic space. In addition, the implicit information inferences are of special interest as they can be viewed as potential realizations of Peirce's "new ideas". In summary, information flow through a HAL space would appear to be an interesting mechanism for driving a logic of discovery motivated from a psychologistic perspective. In the next section, it will be used to replicate Swanson's discovery.

---

[3] Precision is the ratio of relevant retrieved documents to retrieved documents for a given query.

[4] The Text Retrieval Conference (TREC) series is coordinated by the National Institutes of Standards and Technology in Washington (trec.nist.gov). Standard query topics, document collections and relevance judgments allow the performance of document retrieval systems to be benchmarked.

[5] The underlying HAL space was constructed from a collection of 84,678 Associated Press news feeds.

4 Experiment: Replicating Swanson's Raynaud - fish oil discovery

Swanson's explanatory hypothesis was based on intermediate concepts: so-called B-terms. Let A represent "fish oil", and C represent "Raynaud". The implicit connection between A and C was discovered by a set of explicit connections $A - B$ and $B - C$ [24]. Three B-terms were involved: "blood viscosity", "platelet aggregation" and "vascular reactivity". It is important to note that the $A - B$ and $B - C$ connections were present in the respective bodies of scientific knowledge surrounding dietary fish oils and Raynaud's disease. However, the $C \rightarrow A$ connection is hidden as it bridges these disparate bodies of knowledge. Swanson describes this connection as "undiscovered public knowledge" [22].

Information flow is computed using C (Raynaud) as the source concept. It then is examined whether A (fish oil) appears in the ranked list of implicit information inferences. Our assumption was that the intermediate B-terms would be significant carriers of information flow between C (Raynaud) and A (fish oil).

4.1 Data

A corpus of 111,603 MEDLINE core clinic journal articles in the period 1980-1985 were downloaded from the internet[6]. Note that these articles precede the article documenting Swanson's discovery, which also happens to be present in the MEDLINE collection. Swanson made his serendipitous discovery by perusing only document titles, so only these were used to construct a semantic space via HAL. The vocabulary consists of all words in titles except those appearing in a stop word list originating from the ARROWSMITH system [23]. The size of the resulting vocabulary is 28,834 words, which is also the dimensionality n of the HAL space.

4.2 Method

The parameters that were manipulated in the experiment were the size of the window l and the threshold parameter δ. (Recall the latter determines the quality properties of the source concept for information flow computation). Window size was manipulated because we suspected that larger window sizes may be more conducive to capturing B-terms. The threshold parameter was manipulated because we felt that only highly weighted dimensions in the Raynaud representation are relevant to abducing potentially interesting terms. This involves a trade-off: very highly related dimensions may be more conducive for abducing salient terms, but may miss others (i.e., favouring soundness over completeness, so to speak), whereas using many or all of the positively weighted dimensions in the Raynaud representation to drive the inference process may abduce many, or all, of the salient terms among many extraneous ones (i.e., favouring completeness over soundness).

The top 1500 weighted terms were computed by using the "Raynaud" representation as the source concept in Equation 3.1. Implicit information inferences were identified and ranked according to descending order of information flow. The ranking was then inspected for the terms fish, cod, liver and oil. These terms represent the set of salient

[6] PubMed: www.ncbi.nlm.nih.gov/entrez

suggestions.

In order to place information flow computations in the context of other mechanisms used in the literature, other types of association were computed based on the Raynaud representation. Cosine has been used in semantic spaces, most notably those produced by latent semantic analysis [11]. Cosine measures the angle between representations in the semantic space, the smaller the angle, the stronger the association. In the HAL space, all representations are normalised to unit length, so the cosine can be computed by calculating the dot product between respective representations and ranking them on decreasing order of cosine. The top 1500 terms were selected for analysis; the cut-off at 1500 is arbitrary, but reflects our intuition that more than 1500 is "too many" for a logic of justification to handle. The Minkowski distance metrics advocated by Gärdenfors [7] have produced encouraging results in various semantic association experiments using semantic spaces computed by HAL [15, 4]. The distance between two concepts x and y in a n-dimensional HAL space can be calculated using the Minkowski distance metric:

$$d(x,y) \quad = \quad \sqrt[r]{\sum_{i=1}^{n}(|w_{xp_i} - w_{yp_i}|)^r}$$

where $d(x,y)$ denotes the distance between the HAL representations for concepts x and y. Both Euclidean distance ($r = 2$) and city-block distance ($r = 1$) were computed where x corresponds to the HAL representation for the term "Raynaud". The y-terms were ranked on increasing order of distance as terms closer to x are assumed to be more semantically related. The top 1500 y-terms were selected for analysis.

The distinction between information flow computation and cosine/Minkowski should be noted. Cosine and the Minkowski distance metric(s) measure the strength of semantic association between two terms in the space, whereas information flow computation measures the degree of information containment of the target term with respect to the source term.

4.3 Results

The following table depicts the results. Only the best performing runs are reported.

	Cod	Liver	Oil	Fish
Information flow ($l = 50, \delta = \mu$)	**0.12 (484)**	**0.34 (54)**	**0.12 (472)**	0.04
Cosine ($l = 50$)	**0.13 (152)**	0.04	0.04	0.06
Euclidean distance ($l = 50$)	**1.32 (152)**	1.38	1.38	**1.37 (1088)**

TABLE 2: Implicit information inference and semantic association strengths based on the "Raynaud" HAL representation

The numbers in the cells represent the strength of information flow, or semantic association, with the associated ranking in brackets. Values in bold denote terms which appeared in the top 1500 in the respective ranking. Runs based on the city-block metric are not reported due to their poor performance.

4.4 Discussion

It is promising that information flow through a semantic space was able to bridge the implicit connection between "Raynaud" and three out of the four target terms. It is not necessarily discouraging that the rankings of the terms are fairly low (cod: 484, liver: 54, oil 472) because the information flow computations are meant to uncover potentially interesting suggestions - it is the task of a logic of justification to determine which of these, if any, will be entertained.

It is somewhat surprising that the best performing information flow computations were achieved using above average weight quality properties (dimensions) in the Raynaud representation ($\delta = \mu = 0.10$). Inspection of the Raynaud representation revealed that only one B-term was above average weight (platelet: 0.15) whereas other B-terms were present in the representation, but below average weight (platelets: 0.07, viscosity: 0.07, vascular: 0.07, blood: 0.07), so they did not contribute to information flow. Relevant information flow was apparently carried by a single B-term (platelet). It is promising that HAL's lexical co-occurrence does capture B-terms in the representation, but the weighting is not optimal for computing suggestions. In order to gain insight into the potential effectiveness of information flow, we manually increased the weights of the B-terms in the Raynaud vector to maximal value (unity). The resultant vector is denoted "Raynaud+". The second run emulates the case where a researcher manually enhances the automatically derived HAL vector for "Raynaud" motivated from their specific expertise and interest. The run also potentially represents an upper bound on performance as the weights ascribed to the B-terms are maximal. The results are depicted in the table 3. The results are encouraging as

	Cod	Liver	Oil	Fish
Information flow ($l = 50, \delta = \mu$)	**0.31 (273)**	**0.54 (30)**	**0.56 (27)**	**0.59 (17)**
Cosine ($l = 50$)	**0.28 (92)**	**0.07 (1404)**	**0.20 (245)**	**0.20 (245)**
Euclidean distance ($l = 50$)	**1.20 (56)**	**1.36 (823)**	**1.27 (1471)**	2.51

TABLE 3: Implicit information inference from the manually contextualized Raynaud representation "Raynaud+"

significant information flow has been imparted to all target terms, and three of the four target terms have been promoted very high in the ranking. The equivalent experiments using semantic association show improvement over the results in table 2, but the level of improvement is less encouraging than that produced by information flow computation.

The window size ($l = 50$) is much larger than reported in the literature. However, this value essentially means that all terms within an individual title are assumed to be associated for the purposes of creating the semantic space. This is reasonable for title texts, but would not necessarily be applicable if abstracts were used to construct the semantic space.

For reasons of computational efficiency, information flows are computed to individual words. As a consequence, interpreting the ranked output is made difficult. For example, does the term "liver" relate to "cod liver", or the organ? In order to counter this problem, we employed shallow natural language processing to post process the ranked list. It is not relevant to elucidate the details of this process. Suffice to say

that shallow natural language processing allows the ranked list of information flows (i.e., a list of individual terms) to be to be projected into a lattice-like structure which has been derived from parse trees of document titles. Integral dimensions, which are not present in the HAL representations, can be recovered from the lattice structure because it captures syntagmatic relationships (which are lost when applying HAL). After projection into the lattice, the ranked list of information flow terms is rendered into a ranked list of noun compounds like "ventricular infarction", "liver cell carcinoma" and "cod liver oil" (see table 3 in [2]). It is these compounds which would comprise the raw input into a logic of justification.

Euclidean distance (see table 2) also uncovers two of the four target terms (cod: 1.32, fish 1.37). It seems that semantic association may form the basis of some useful suggestions. Observe that "fish" was detected via semantic association, but was not detected by information flow computation, whereas information flow uncovered "liver" and "oil" both of which were undetected by semantic association. Therefore, on their own, both information flow and semantic association computed via Euclidean distance do not offer a complete mechanism for automatically producing hypotheses.

5 Conclusions and further research

This article provides an account of how implicit connections can be computed from a semantic space and interpreted from an abductive perspective. In an automatic setting, information flow computation through a high dimensional semantic space is able to suggest the majority of terms needed to simulate Swanson's Raynaud-fish oil discovery, though the strength of suggestion is relatively small. Semantic representations of concepts derived from the Hyperspace Analogue to Language method seem to capture many relevant associations to salient intermediate terms, however, these terms are not weighted optimally. As a consequence, the ability to compute relevant abductions from the semantic space is impinged. Further research is needed to develop an appropriate contextualizing function which promotes the weight of salient intermediate terms in the underlying semantic representation.

In a semi-automated setting, the strength of suggestion of relevant terms is promoted. This is encouraging as more strongly weighted suggestions are those which will be considered by a logic of justification, the goal of which is to filter the hypotheses into a small set of candidates for perusal by a human. It should be noted that the Raynaud-fish oil connection is indirect and other fully automated attempts have often been disappointing, and semi-automated attempts have had mixed success in detecting it [9, 8, 24, 21]. In short, we feel that there is some justification to be believe that HAL-based information flow, perhaps enhanced by suggestions computed by other means (e.g., semantic association, inference by dimensional reduction), can underpin a logic of discovery for textual information. The goal of such a logic is simply to produce potentially interesting suggestions, or hypotheses, which can span disparate islands of knowledge. Future work will be directed toward realizing a logic of justification. Thereafter we plan to replicate other discoveries, such as Swanson's migraine-magnesium hypothesis [23].

Our conviction is that a logic of discovery should not be conceived in a traditional sense - there are no rules of inference which prescribe its behaviour, but rather suggestions are produced by computing implicit associations within a semantic (geometric)

space whereby the knowledge representation is motivated from a cognitive perspective. We feel that this view of a "logic" is aligned with C.S.Peirce's view of abduction: "No reason whatsoever can be given for it [abduction], as far as I can discover; and it needs no reason, since it merely offers suggestions" ([17], p217).

Acknowledgments

The work reported in this paper has been funded in part by the Co-operative Research Centre for Enterprise Distributed Systems Technology (DSTC) through the Australian Federal Government's CRC Programme (Department of Industry, Science & Resources). We gratefully thank Marc Weeber for supplying the stopword list from the ARROWSMITH system.

References

[1] J. Barwise and J. Seligman. *Information flow: the logic of distributed systems.* Cambridge University Press, 1997.

[2] P.D. Bruza, R.M. McArthur, and D. Song. Discovery of explicit and implicit connections in textual information, 2003. Online: http://www.dstc.edu.au/Research/Projects/Infoeco/publications.

[3] P.D. Bruza and D. Song. Inferring Query Models by Computing Information Flow. In *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM 2002)*, pages 260–269. ACM Press, 2002.

[4] C. Burgess, K. Livesay, and K. Lund. Explorations in context space: words, sentences, discourse. *Discourse Processes*, 25(2&3):211–257, 1998.

[5] D. Gabbay and J. Woods. *Agenda Relevance: A Study in Formal Pragmatics*, volume 1 of *A Practical Logic of Cognitive Systems.* Elsevier, 2003.

[6] D. Gabbay and J. Woods. *The Reach of Abduction: Insight and Trial*, volume 2 of *A Practical Logic of Cognitive Systems.* Elsevier, 2004. An early draft appeared as Lecture Notes from ESSLLI 2000 (European Summer School on Logic, Language and Information), Online: http://www.cs.bham.ac.uk/ esslli/notes/gabbay.html.

[7] P. Gärdenfors. *Conceptual Spaces: The Geometry of Thought.* MIT Press, 2000.

[8] M.D. Gordon and S. Dumais. Using Latent Semantic Indexing for literature-based discovery. *Journal of the American Society for Information Science*, 49(8):674–685, 1998.

[9] M.D. Gordon and S. Lindsay. Toward Discovery Support Systems: A Replication, Re-Examination, and Extension of Swanson's Work on Literature-Based Discovery of a Connection between Raynaud's and Fish Oil. *Journal of the American Society for Information Science*, 47(3):116–128, 1996.

[10] 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*, 104:211–240, 1997.

[11] T.K. Landauer, P.W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25(2&3):259–284, 1998.

[12] J.P. Levy and J.A. Bullinaria. Learning lexical properties from word usage patterns: Which context words should be used? In R.F. French and J.P. Sounge, editors, *Connectionist Models of Learning, development and Evolution: Proceedings of the Sixth Neural Computation and psychology Workshop*, pages 273–282. Springer, 1999.

[13] W. Lowe. What is the dimensionality of human semantic space? In *Proceedings of the 6th Neural Computation and Psychology workshop*, pages 303–311. Springer Verlag, 2000.

[14] W. Lowe. Towards a theory of semantic space. In J. D. Moore and K. Stenning, editors, *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, pages 576–581. Lawrence Erlbaum Associates, 2001.

[15] K. Lund and C. Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers*, 28(2):203–208, 1996.

[16] M. Patel, J.A. Bullinaria, and J.P. Levy. Extracting semantic representations from large text corpora. In R.F. French and J.P. Sounge, editors, *Connectionist Models of Learning, Development and Evolution: Proceedings of the Fourth Neural Computation and Psychology Workshop*, pages 199–212. Springer, 1997.

[17] C.S Peirce. The Nature of Meaning. In Peirce Edition Project, editor, *Essential Peirce: Selected Philosophical Writings Vol 2 (1893-1913)*, pages 208–225. Indiana Univ. Press, 1998.

[18] M. Sahlgren. Towards a Flexible Model of Word Meaning. Paper presented at the AAAI Spring Symposium 2002, March 25-27, Stanford University, Palo Alto, California, USA, 2002.

[19] D.W. Song and P.D. Bruza. Discovering Information Flow using a High Dimensional Conceptual Space. In *Proceedings of the 24th Annual ACM Conference of Research and Development in Information Retrieval (SIGIR'2001)*, pages 327–333. ACM Press, 2001.

[20] D.W. Song and P.D. Bruza. Towards context sensitive information inference. *Journal of the American Society for Information Science and Tecnology*, 54(3):321–334, 2003.

[21] P. Srinivasan. Text Mining: Generating Hyptheses from MEDLINE. *Journal of the American Society for Information Science and Technology*, 55(5):396–413, 2004.

[22] D.R. Swanson. Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18, 1986.

[23] D.R. Swanson and N.R. Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91(2):183–203, 1997.

[24] M. Weeber, H. Klein, L. Jong van den Berg, and R. Vos. Using concepts in literature-based discovery: Simulating Swanson's Raynaud-Fish Oil and migraine-magnesium discoveries. *Journal of the American Society for Information Science and Technology*, 52(7):548–557, 2001.