Towards Context Sensitive Information Inference

D. Song and P.D. Bruza
Distributed Systems Technology Centre
The University of Queensland, Australia
{dsong|bruza}@dstc.edu.au

Abstract

Humans can make hasty, but generally robust judgements about what a text fragment is, or is not, about. Such judgements are termed information inference. By drawing on theories from non-classical logic and applied cognition, an information inference mechanism is proposed which makes inferences via computations of information flow through a high dimensional conceptual space. Within a conceptual space information is represented geometrically. In this article, an approximation of a conceptual space is employed whereby geometric representations of words are realized as vectors in a high dimensional semantic space, which is automatically constructed from a text corpus. Two approaches were presented for priming vector representations according to context. The first approach uses a concept combination heuristic to adjust the vector representation of a concept in the light of the representation of another concept. The second approach computes a prototypical concept on the basis of exemplar trace texts and moves it in the dimensional space according to the context. Information inference is evaluated by measuring the effectiveness of query models derived by information flow computations. Results show that information flow contributes significantly to query model effectiveness, particularly with respect to precision. Moreover, retrieval effectiveness compares favourably with two probabilistic query models, and another based on semantic association. More generally, this article can be seen as a contribution towards realizing operational systems which mimic human text-based reasoning.

1. Introduction

Information is like a field with ever receding boundaries. Paradoxically, the expansion of this field is leading to a diminishing awareness. The reason for this is grounded in the fact that human beings have limited resources, chiefly time and cognitive processing power. The information field increases more swiftly than the corresponding growth in the human being's resources. As a consequence, disciplines and expertise are becoming increasingly specialized with little awareness of kindred specializations. Swanson's (1997) serendipitous literature-based discovery of a cure for Raynaud's disease by dietary fish oils illustrates this phenomenon. The literature documenting Raynaud's disease and literature surrounding fish oil were disjoint. 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." (Swanson, 1997 p184). If these communities would have been aware of each other, a cure would probably been found much earlier than Swanson's discovery.

An important limiting factor in our ability to process information is our bounded cognitive firepower. We simply cannot absorb information as quickly as it is growing. It is interesting to tarry briefly by the question of why this is the case. Information-theoretic research into consciousness has revealed that it has a surprisingly narrow bandwidth. It processes information slowly. The rate of processing of our senses is in the region of between 10^7 and 10^{11} bits per second. For any given second, only about 16 to 20 bits of information enter into consciousness (Austin, 1998 p278). Consciousness, therefore, "is highly entropic, a thermodynamically costly state for the human system to be in" (Gabbay & Woods, 2000 p67). It would seem that human beings make do with little information because that is all the conscious individual can handle.

Gabbay and Woods (2000) have recently proffered a notion of cognitive economics founded on compensation strategies employed by the human agent to alleviate the consequences of its limited resources.

We briefly recount aspects these compensation strategies and then dovetail them into a discussion specifically centred on the cognitive economics surrounding textual information processing.

One compensation strategy employed by agents is to divide reality into natural kinds, e.g., feathered, flying objects are "birds". Paired with natural kinds is hasty generalization, also known as generic inference[1]. (Gärdenfors, 2000). By way of illustration, after seeing a feathered, living object fly past, the human agent may leap to the conclusion that "birds fly". Such generalizations are based on small sample sizes of natural kinds, and are defeasible in the light of new experience: On encountering an emu the previous generalization will be defeated, and perhaps lead to an adjustment of the associated natural kind into "typical bird", and "atypical bird", the former of which would still support the generalization. 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 (2000 p64) observe, “If generic inferences from natural kind samples are note *quite* right, at least they don’t kill us. They don’t even keep us from a certain abundance in life”.

Cognitive Economics of Information Processing

Due to limited cognitive resources, human agents tend to accept information from other agents, because, by doing so, the agent does not have to perform the processing itself. In other words, acceptance of information from another party is an economic way of acquiring information by default. Three principles in this regard are mentioned in passing below. For a detailed exposition of these principles in relation to human (abductive) reasoning, the reader is referred to (Gabbay & Woods, 2000):

- *Ad Verecundium* (appeal to authority): acceptance of reputed opinion.
- *Ad populum*: acceptance of popular belief (*endoxa*)
- *Ad Ignorantiam*: Human agents tend to accept, without challenge, the arguments of others except where they know, or think they know, or suspect, that something is amiss.

When helpful other agents are not present, the agent is forced to digest the information on its own. A compensation strategy we dub *Ad compactium* alludes to the preference of the agent for compact, easily digestible, information chunks. (How often does one hear the call for a good, succinct introductory text on a certain topic?)

A dramatic compensation strategy rejects information. Rejection is an economic strategy in the extreme, as rejected information need not be processed at all. It is reasonable to assume that agents resort to this strategy when "stressed", for example, when dealing with a huge influx of e-mail, any excuse will be used to delete a given e-mail, or file it away "for future processing".

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 judgements in regard to what the information is, or is not, about. Such judgements are 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 readily 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 judgement 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.

[1] It is also known as inductive inference (Gärdenfors, 2000)

In short, human beings can generally make robust judgements about what information fragments are, or are not about, even when the fragments are brief, or incomplete. The process of making such judgements will be referred to as *information inference*.

Inference and information

The above examples attempt to illustrate that information inference is a very real phenomenom. 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 resources. Barwise and Seligman (1997) 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 the background knowledge *k* includes "light bulbs need electricity to burn, flicking the switch on allows electricity to flow, etc". It is instructive to see how the above definition functions with respect to the information inferences drawn from the text fragments of the previous section.

Information inferences are sometimes 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"*[2]. Analogously, *"Penguin" being (together with) "Books"* carries the information that *"Penguin" is publisher.*

On the basis of these examples, it would seem that 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 inferences[3]. 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" (Barwise & Seligman, 1997 p23). This position poses an immediate and difficult challenge due to the inherent flexibility required. In our case, we will 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 the 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" (Barwise

[2] The fact that *"Linus" being with "penguin"* strengthens this inference

[3] Seligman is a logician with a decidedly non-classical bent. Before his untimely death, the same can be said about Barwise. Their overtly psychologistic stance is also adopted by Gabbay and Woods (2000). The classical tradition of the last century seems in the process of being challenged by the "New Logic", which, at the very least, appears willing to seriously consider psychologism. See Woods and Gabbay (2000) for a historical perspective.

& Seligman, 1997 p23). Again this poses a challenge to implementing an information inference system. How will *k* be acquired, used appropriately, and kept up-to-date?

Barwise and Seligman's definition stands in contrast to probabilistic definitions of information inference, the most prominent of which originates from Dretske (Barwise & Seligman, 1997 p15):

Dretske's Information Content: To a person with prior knowledge *k*, *r being F* carries the information that *s is G*, if and only if the conditional probability of *s being G* given that *r is F* is 1 (and less than one given *k* alone)

Barwise and Seligman object to Dretske's definition on two counts. The first objection centers around Dretske's requirement that the conditional probability must be one in order for information inference to occur. Barwise and Seligman consider that "it sets too high a standard". For example, snowboarding is analogous to surfing. The probability that "Surfing the Himalayas" is about snowboarding, given this fact, may be high, but is not necessarily certain, Dretske's definition won't permit it. A human being, however, may leap to the conclusion that it is about snowboarding.

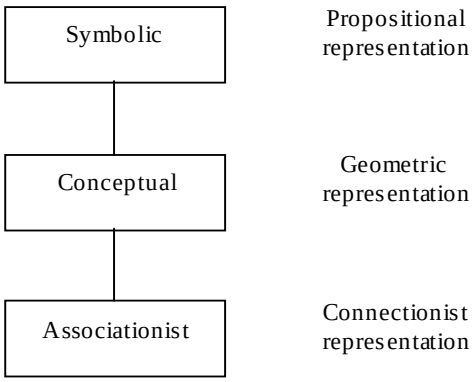
A section objection is that Dretske's definition "makes no room for *a priori*" knowledge. Consider the statement, *s* = "the earth being a planet". It would seem reasonable that this statement carries the information that "the earth is round". Observe that the probability of *s* is one. Then, given any prior knowledge *k*, the probability of *s* given *k* is also one, so, according to Dretske's definition, the desired information inference cannot proceed.

We accept Barwise and Seligman's rebuttal of Dretske's approach. Moreover, we reject Dretske's definition out of principle, as it is incompatible with our psychologistic stance.

This article will attempt to furnish a practical account of information inference. In a nutshell, it will attempt to realize Barwise and Seligman's inferential information content definition by computing information flows within a high dimensional conceptual space which is motivated from a cognitive perspective. The next section introduces conceptual spaces in terms of a three level model of cognition. Section 3 describes how an operational approximation to a conceptual space may be derived from a text corpus. Section 4 details how information inference may be realized via information flow computations through the conceptual space. Practical investigations into information inference are reported in section 5 where it is applied to the derivation of query models for information retrieval. The article is rounded off with a summary and conclusions.

2. Conceptual Space

Gärdenfors (2000) has recently proposed a model of cognition whereby symbolic processing is but one of three levels. These levels can be seen "three levels of representation of cognition with different scales of resolution". How information is represented varies greatly across the different levels:



Within the associationist level, information elements are connected via associations, for example, connectionism. Connectionist systems consist of highly interconnected nodes (neurons), which process information in parallel. Gärdenfors deems the associationist level to be "sub-conceptual". This level will not play a role in this article, so it will not be discussed further.

At the conceptual level, information is represented geometrically in terms of a dimensional space. For example, the property colour can be represented in terms of three dimensions: hue, chromaticity, and brightness. Hue is manifested directly from the wavelength of the light. Chromaticity is a dimension that reflects the saturation of the colour. Brightness can be represented by a dimension with values ranging from 0 (dark) to 1 (bright). Gärdenfors argues that a property is represented as a convex region in a geometric space. In terms of the example, the property "red" is a region within the tri-dimensional space made up of hue, chromaticity and brightness. The property "blue" would occupy a different region of this space. The three dimensions that together represent the property of colour have their roots in the human perceptual mechanism, however, this need not always be the case; dimensions may also be abstract.

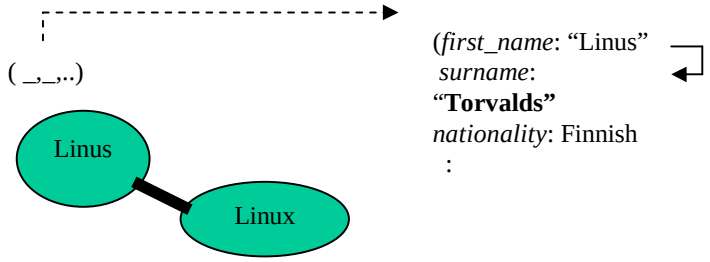
Gärdenfors extends the notion of properties into concepts which are based on domains. A domain is a set of integral dimensions in the sense that a value in one dimension(s) determines or affects the value in another dimension(s). For example, the three colour dimensions are integral as the brightness of a colour will affect its saturation (chromaticity) and hue.

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). Gärdenfors argues that concepts can be learnt from a limited number of exemplars of the concept. The prototype of the concept is assumed to be a "typical" instance, which is extracted from the exemplars. For example, one may hold that the typical apple is "red". Exemplars are described by points in the conceptual space. One way of calculating the prototype from a set of exemplars is that the *i*-th coordinate p_i for the vector *p* representing the prototype is defined to be the mean of the *i*-th coordinate of all the exemplars.

Observe the distinction between representations at the symbolic and conceptual levels. At the symbolic level "apple" could be modelled as the atomic proposition *apple(x)*, however, at the 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 propositional 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".

Inference at a conceptual level is brought about relations between concepts, or triggered associations of a concept within a given context. Both "Linux" and "Linus" are concepts consisting of a number of domains. The text "Linux Online: Why Linus chose a Penguin" triggers the association between the concept "Linux" with the concept "Linus". In addition, it disambiguates: Linus the inventor of Linux, as opposed to Linus the

cartoon character from “Peanuts”. As a consequence certain domains relevant to the inventor of Linux are weighted prominently, for example, first name, surname, nationality etc. First-name and surname are integral so the value “Linus” in the first-name dimension is inter-related with the value “Torvalds” in the surname dimension. This association ultimately leads to the inference that the text deals with “Linus Torvalds”.



The above example is a simplistic illustration of Gärdenfors’ conceptual spaces. In addition, it cannot totally verified from a cognitive perspective. For example, research in cognitive science has not clarified whether Linus is a single concept from which salient domains are drawn according to the context, or whether there are two distinct concepts of Linus – the one corresponding to Linus Torvalds, and another corresponding to Linus, the cartoon character. Nevertheless, the appeal of Gärdenfors’ conceptual spaces is that it allows inference to be considered not only at the symbolic level, but at the conceptual (geometric) level as well. Observe that inference at the symbolic level is typically modelled by a liner sequence of propositions prescribed by rules of inference. Within the conceptual level, inference takes on a decidedly associational character. This is not only interesting from a cognitive point of view, but also opens the door to computationally tractable inference mechanisms. 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” (Gärdenfors, 2000 p127).

Barwise and Seligman (1997) also propose a geometric foundation to their account of inferential information content via the use of real-valued state spaces. The basic assumption is that the state of any system can be determined by the values taken by various attributes, the observables of the system. (An observable corresponds to Gärdenfors’ notion of dimension). Observables are measurable quantities, the associated value being a real number. An example of an observable is temperature. A state of a σ system of n observables is represented by a n-dimensional vector of reals, $\sigma \in \Re^n$. Observe that this is a similar proposal, albeit more primitive, to that of Gärdenfors with respect to properties in the conceptual space. Moreover, Barwise and Seligman also account for integral dimensions in the underlying vector representation by the use of observation functions. These functions prescribe how the values in certain dimensions determine the value in another dimension. At this point, we will not delve further into the technical details of Barwise and Seligman’s real-valued state spaces. There relevance to our account is the following. Firstly, real-valued state spaces open the door to the practical realization of information inference by implementing a conceptual space appropriately as a real-valued state space. This is the subject of the next section. Secondly, real-valued state spaces have a fundamental role to play with respect to our psychologistic point of departure. We firmly agree with Barwise and Seligman who state, “Within the recent cognitive-science literature, logic is often seen as irrevocably wed to what is perceived to be an out-dated symbol processing model of cognition. From there it is but a short step to the conclusion that the study of logic is irrelevant for cognitive science. This step is often taken in spite of the fact that human reasoning is a cognitive activity an do must be part of cognitive science. *Perhaps the use of state spaces might allow a marriage of logic with continuous methods like those used in dynamical systems*

and so provide a toehold for those who envision a distinctively different model of human reasoning" (Barwise and Seligman, 1997 p234 - italics ours)

3. Towards an Implementation of Conceptual Spaces: Hyperspace Analogue to Language

A human encountering a new concept derives the meaning via an accumulation of experience of the contexts in which the concept appears. This opens the door to "learn" the meaning of a concept through how a concept appears within the context of other concepts. Following this idea, a representational model of semantic memory has been developed called Hyperspace Analogue to Language (HAL), which automatically constructs a dimensional semantic space from a corpus of text (Burgess et al, 1998; Burgess & Lund, 1997; Lund & Burgess, 1996). The space comprises high dimensional vector representations for each term in the vocabulary. Given an *n*-word vocabulary, the HAL space is a *n* x *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. After traversing the corpus, an accumulated co-occurrence matrix for all the words in a target vocabulary is produced. HAL is direction sensitive: the co-occurrence information for words preceding every word and co-occurrence information for words following it are recorded separately by its row and column vectors. By way of illustration, the HAL space for the example text "The effects of spreading pollution on the population of Atlantic salmon" is depicted below (Table 1) using a 5 word moving window (*l*=5):

	the	eff	of	spr	poll	on	Pop	atl	sal
the		1	2	3	4	5			
eff	5								
of	8	5		1	2	3	5		
spr	3	4	5						
poll	2	3	4	5					
on	1	2	3	4	5				
pop	5		1	2	3	4			
atl	3		5		1	2	4		
sal	2		4			1	3	5	

Table 1: Example of a HAL space

This table shows how the row vectors encode preceding word order and the column vectors encode posterior word order. For the purposes of our account of information inference, we deem it unnecessary to preserve order information, so the HAL vector of a word is represented by the addition of its row and column vectors. As an example of a HAL vector derived from a large corpus, consider part of the normalized HAL vector for "*superconductors"* computed from a corpus of Associated Press news:

superconductors = < U.S.:0.11 american:0.07 basic:0.11 bulk:0.13 called:0.15 capacity:0.08 carry:0.15 ceramic:0.11 commercial:0.15 consortium:0.18 cooled:0.06 current:0.10 develop:0.12 dover:0.06 electricity:0.18 energy:0.07 field:0.06 goal:0.06 high:0.34 higher:0.06 improved:0.06 japan:0.14 loss:0.13 low:0.06 make:0.07 materials:0.25 new:0.24 require:0.09 research:0.12 researching:0.13 resistance:0.13 retain:0.06 scientists:0.11 semiconductors:0.10 states:0.11 switzerland:0.06 technology:0.06 temperature:0.48 theory:0.06 united:0.10 university:0.06>

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 "superconductors" 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 "superconductors" 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. Burgess *et al*. (1998) use a window size of eight or ten in their studies, though the motivation for these numbers is not compelling (Perfetti, 1998).

Burgess *et al*. (1998) were able to demonstrate the cognitive compatibility of HAL vectors with human processing via a series of word matching and word similarity experiments. We therefore feel that HAL space is a promising candidate on which to found and information inference mechanism whereby the inferences would correlate with those a human would make. In addition, a HAL space is a real-valued state space, whereby each vector can be considered to represent the "state" of a particular word in relation to the corpus from which the HAL space was constructed. More formally, a concept[4] c is a vector representation: $c = < w_{cp_1}, w_{cp_2}, ..., w_{cp_n} >$ where $p_1, p_2, ..., p_n$ are called dimensions of *c*, *n* is the dimensionality of the HAL space, and w_{cp_i} denotes the weight of p_i in the vector representation of *c*. A dimension is termed a property if its weight is greater than zero. A property p_i of a concept c is a termed a *quality property* iff $w_{cp_i} > \partial$, where ∂ is a non-zero threshold value. Let $QP_\partial(c)$ denote the set of quality properties of concept *c*. $QP_\mu(c)$ will be used to denote the set of quality properties above mean value, and $QP(c)$ is short for $QP_0(c)$.

In summary, the HAL vector for a concept *c* is a vector whose non-zero dimensions represent words co-occurring with *c* in the context of a window somewhere in the corpus. The weights in these dimensions represent how strongly *c* has been seen in the context of other words within the corpus used to define the HAL space. An objection that can be levelled at such a global lexical co-occurrence approach is context-insensitivity. For example, the vector representation for "penguin" may have dimensions relating both to the animal and publisher sense of the word. The problem arises as how to weight the dimensions appropriately according to the context. For example, in the publisher context, the weights in dimensions of the concept penguin relating to the animal should be drastically reduced. The question is as to how such "contextualizing" of the vector can be brought to bear. As mentioned previously, Gärdenfors (2000) advocates a weighting function across the dimensions which weights the dimensions appropriately according to the context. Such a function is an apt means for *representing* context and *applying* its effects, but the question beckons as to how this function can be *primed* in practice. Consider once again the example text "Penguin Books U.K.". Neighbouring words to "Penguin" give clues as to what dimensions of the underlying vector should be emphasized. In the following section, we propose a heuristic that contextualizes the penguin vector by combining it with vectors of other concepts in its neighbourhood.

Context sensitivity via Concept Combination

Our ability to combine concepts and, in particular, to *understand* new combinations of concepts is a remarkable feature of human thinking, for example, "space program" and "pink elephant". Within a conceptual space, the combination of concepts can be realized in terms of the respective geometric representations. For example, one domain of the concept "elephant" is colour, which is typically grey. The concept "pink elephant" can be constructed by replacing this grey region with another region representing the property pink. Observe in this example that "pink" acts as a modifier for the concept "elephant". In general, the combination of concepts cannot always be realized in such a straightforward fashion. For example, the geometric representation of "space program" would include domains from both concepts, but it is not obvious how the resulting geometric representation could be realized. It is reasonable to assume that the resultant concept would be more "space-ish" rather than "program-ish", possibly because "space" is a more specific term. We refer to this intuition as *dominance* - the concept "space" is said to dominate the concept "program" in the formation of the associated concept combination.

[4] The term "concept" is used somewhat loosely to emphasize that a HAL space is a primitive realization of a conceptual space

Gärdenfors introduces a concept combination rule, which offers the beginnings of a computational procedure. However, this rule is defined in terms of regions and so-called contrast classes, which are not available in the more primitive real-valued state spaces. For this reason, we have developed a heuristic concept combination specifically for HAL spaces.

Given two concepts $c_1 = \langle\ w_{c_1p_1}, w_{c_1p_2}, ...w_{c_1p_n}\ \rangle$ and $c_2 = \langle\ w_{c_2p_1}, w_{c_2p_2}, ...w_{c_2p_n}\ \rangle$, whereby concept c_1 is assumed to dominate concept c_2. The resulting combined concept is denoted $c_1 \oplus c_2$. The heuristic is essentially a restricted form of vector addition whereby quality properties shared by both concepts are emphasized, the weights of the properties in the dominant concept are re-scaled higher, and the resulting vector from the combination heuristic is normalized to smooth out variations due to differing number of contexts the respective concepts appear in.

Step 1: Re-weight c_1 and c_2 in order to assign higher weights to the properties in c_1.

$$w_{c_1p_i} = \ell_1 + \frac{\ell_1 * w_{c_1p_i}}{\underset{k}{Max}(w_{c_1p_k})}$$

$$w_{c_2p_i} = \ell_2 + \frac{\ell_2 * w_{c_2p_i}}{\underset{k}{Max}(w_{c_2p_k})}$$

$\ell_1,\ \ell_2 \in (0.0,\ 1.0]$ and $\ell_1 > \ell_2$

This step enforces dominance by re-scaling the weights. The dimension weights in the respective concepts are scaled with respect to the parameters l_1 (for concept c_1) and l_2 (for concept c_2). For example, if $\ell_1 = 0.5$ and $\ell_2 = 0.4$, then property weights of c_1 are transferred to interval [0.5, 1.0] and property weights of c_2 are transferred to interval [0.4, 0.8], thus scaling the dimensions of the dominant concept higher.

Step 2: Strengthen the weights of quality properties appearing in both c_1 and c_2 via a multiplier α; the resultant highly weighted dimensions constitute significant properties in the resultant combination.

$\forall(p_i \in QP_{\partial_1}(c_1) \wedge p_i \in QP_{\partial_2}(c_2))$ [$w_{c_1p_i} = \alpha * w_{c_1p_i}$ *and* $w_{c_2p_i} = \alpha * w_{c2p_i}$] where $\alpha > 1.0$

Step 3: Compute property weights in the composition $c_1 \oplus c_2$ via vector addition:

$$w_{(c_1 \oplus c_2)p_i} = w_{c_1p_i} + w_{c_2p_i}$$

Step 4: Normalize the vector $c_1 \oplus c_2$. The resultant vector can then be considered as a new concept, which, in turn, can be composed to other concepts by applying the same heuristic.

The above heuristic depends on five tuning parameters:

- l_1 and l_2 determine how dominance is reflected in dimension weights.
- ∂_1 and ∂_2 determine which quality properties of the respective concepts will have their values strengthened by the multiplier α. These three parameters are used to determine what dimensions will be emphasized in the resultant vector representing the combination.

The above heuristic has a definite ad hoc character, but nevertheless seems to produce desirable representations of combined concepts (Song & Bruza, 2001). For example, the following two normalized HAL vector fragments for the concepts "Reagan" and "Iran" have been derived from applying HAL to the Reuters-21578 collection[5] with parameters (.

Reagan = < administration: 0.46, bill: 0.07, budget: 0.08, congress: 0.07, economic: 0.05, house: 0.09, officials: 0.05, president: 0.80, reagan: 0.09, senate: 0.05, tax: 0.06, trade: 0.09, veto: 0.08, white: 0.06, …>

Iran = < arms: 0.71, attack: 0.18, gulf: 0.21, iran: 0.33, iraq: 0.31, missiles: 0.11, offensive: 0.13, oil: 0.18, reagan: 0.10, sales: 0.20, scandal: 0.25, war: 0.20, … >

The dimensions reflect aspects relevant to the respective concepts during the mid to late eighties. For example, Iran was involved in a war with Iraq, Ronald Reagan was president and he was embroiled in an arms scandal involving Iran and the Contra rebels. The following vector fragment represents the concept combination of "Iran" and "Reagan", assuming "Iran" to be dominant ($\ell_1 = 0.5, l_2 = 0.3, \alpha = 2.0, \partial_1 = 0, \partial_2 = 0$) :

Reagan ⊕ Iran = < administration: 0.11, affair: 0.06, arms: 0.72, attack: 0.08, contra: 0.14, deal: 0.08, diversion: 0.07, gulf: 0.11, house: 0.10, initiative: 0.06, iran: 0.22, november: 0.06, policy: 0.07, president: 0.26, profits: 0.08, reagan: 0.23, sales: 0.15, scandal: 0.31, secret: 0.06, senate: 0.06, war: 0.12, … >

One way of viewing this vector is as a representation of the concept "Iran" (the dominant concept) in the light of the concept "Reagan". Observe how the weights of some dimensions have changed appropriately with respect to associations relevant to Reagan in the context of Iran. More specifically, "arms", "scandal", "contra", are highly weighted, or have had their weights increased, whereas dimensions dealing with Reagan in the general presidential sense have decreased rather dramatically (e.g. "administration"). This illustrates desirable nonmonotonic behaviour with respect to context (Gärdenfors, 2000 p126).

Context sensitivity by "moving" prototypical concepts

As detailed previously, a HAL vector for a concept *c* is constructed by sliding a fixed window over the corpus and accruing weights for the words co-occurring with *c* in the context of the window. The weights in the dimensions represent how strongly *c* has been seen in the context of another words throughout the whole collection.

Another approach to represent *c* is to compute the prototypical concept vector for *c*. By way of illustration, the following are example traces from the Reuters-21578 collection including the term "Reagan":

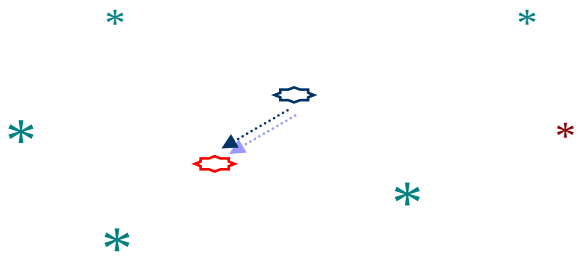
- *"President Reagan was ignorant about much of the Iran arms deal"*
- *"Reagan approval rating falls to four-year low"*
- *"Wallis is quoted as saying the Reagan administration wants Japanese cooperation to ensure the trade bill is a moderate one"*

These traces represent different contexts in which the concept "Reagan" is involved: the Iran-contra arms scandal, politics and trade. Each of the traces above can be used as an "exemplar" of the concept "Reagan", and an associated vector can be constructed by assigning positive weights to those terms which co-occur with Reagan in a given trace, e.g., the same weighting scheme employed by the HAL model can be used. The prototypical Reagan concept can be computed according to Gärdenfors' proposal: form the average vector from the exemplars. The following vector depicts, in decreasing order of weight, the more highly weighted dimensions of the prototypical concept "Reagan" computed from the Reuters-21578 collection using traces of 15 words in length containing the word "Reagan":

[5] The vocabulary was constructed by removing stop words and also dropping some infrequent words which appears less than 25 times in the collection. Window size used: $l = 6$.

Reagan = <president:4.79, administration:2.89, trade:0.95, house:0.90, budget:0.69, congress:0.64, bill:0.59, reagan:0.56, white:0.51, dlrs:0.50, year:0.49, veto:0.49, billion:0.49, japan:0.48, told:0.47, officials:0.46, tariffs:0.44, senate:0.44, economic:0.42, tax:0.42, …>

Note there are dimensions relating to trade, politics, economics, etc. Context-sensitivity can be realized by constructing the prototypical concept based on a *subset* of traces. For example, consider the concept "Reagan" in the context of "Iran". One would expect dimensions dealing with the Iran-Contra scandal to be highly weighted. This can be realized by computing the prototypical concept for "Reagan" based only on those traces mentioning "Iran". This has the effect of shifting the prototype based on a subset of exemplars, as depicted in the figure below. The stars represent points in the high dimensional space corresponding to the exemplar. The flat , large stars correspond to the average of these exemplars (i.e., the prototype). When a subset of exemplars are used, the prototype shifts in the dimensional space:



For example, highly weighted dimensions of the prototypical concept for "Reagan" constructed from traces containing the term "Iran" is as follows:

Reagan = <president:3.82, iran:2.02, **arms**:1.50, administration:1.23, **scandal**:0.74, house:0.73, made:0.72, told:0.67, **speech**:0.64, senate:0.60, **contra**:0.58, conference:0.54, pct:0.53, gulf:0.53, news:0.51, **approval**:0.51, officials:0.50, white:0.49, security:0.46, kuwaiti:0.44, **rating**:0.43, **diversion**:0.42, november:0.42, national:0.42, **tower**:0.42, oil:0.41, reagan:0.40, congress:0.40, **approved**:0.37, **televised**:0.35, ronald:0.35, **poindexter**:0.34, … >

This reveals how dimensions dealing with the Iran-Contra scandal (highlighted with bold) have been promoted. (Poindexter was head of the CIA and the Tower commission report broke the scandal during November. Reagan gave a speech. His approval rating was affected.). In other words, desirable contextualization of the Reagan vector has been achieved. It is interesting to observe that context is not reflexive. Consider the dual of the previous vector: the prototypical concept "Iran" in the context of "Reagan". This vector features dimensions relevant to the diversion of arms to the Contra rebels, whereas the contextualized "Reagan" prototype also includes dimensions relevant to Reagan's response to the scandal (e.g., approval ratings and a televised speech etc.).

Iran = <**arms**:6.80, **scandal**:3.02, reagan:2.85, **sales**:1.71, **contra**:1.63, president:1.23, **profits**:0.84, **secret**:0.81, **rebels:**0.69, gulf:0.69, **sale**:0.67, **diversion**:0.66, **affair**:0.60, friday:0.59, senate:0.59, deal:0.57, policy:0.56, initiative:0.56, **diverted**:0.56, attack:0.54, house:0.54, commission:0.53, report:0.52, made:0.52, iran:0.50, silkworm:0.48, **tower**:0.48, iraq:0.46, **november**:0.45, …>

This section has featured Hyperspace Analogue to Language as a means of deriving a real-valued conceptual space, (a HAL space), wherein concepts and combinations of concepts are represented as high dimensional vectors. Context is considered to be an issue affected by a concept being together with other concepts, , "Reagan" in the context of "Iran", "Linus" in the context of "Linux" etc. In other words, the vector representation of a concept should be conditioned by the presence of other concepts. Such conditioning is referred to as contextualizing the underlying vector representation.

In short, the HAL space offers a computationally tractable information representation mechanism with cognitive credentials whereby context effects can be catered for. The question now is what can be inferred from it.

4. Information Inference via Information Flow

Barwise and Seligman (1997, p54) have proposed an account of *information flow* in terms of state spaces which is a realization of their definition of inferential information content given in the introduction.

Definition 1 (Barwise-Seligman's Information Flow)

$$i_1, \ldots, i_n \mid\!- j \quad \text{iff} \quad \bigcap_{0 \le k \le n} s(i_k) \subseteq s(j)$$

The left hand side of the formula describes a relationship between a set of types (tokens) i_1, i_2, … i_n and a type (token) j. The intuition is the information described by the combination of tokens i_1 to i_n carries the information described by j, for example, $on, live \mid\!- light$ denotes the "switch being on with a live light bulb carries the information that the light is on". Barwise and Seligman refer to this phenomenom as a "constraint" between the respective sets of types; the relationship can be conceived as one of information flow between the conjunction of i_1 to i_n to j. The above definition can also be applied to the earlier examples of information inference, for example, $\text{penguin,books} \mid\!- \text{publisher}$.

The right hand side of Barwise and Seligman's definition describe how the inference relationship is defined in terms of an underlying state space. It is beyond the scope of the article to reproduce the technical details, so we attempt to capture some of the intuition be means of the light bulb example: *s(on)* represents the set of states where the switch is on, *s(live)* represents the set of states where the light bulb is live (i.e., not broken), and *s(light)* represents the set of states where the light is on. The right hand side of the definition equates to those states in which the switch is on and the bulb is live, also entails the light to be on in these states. When dealing with information inference in relation to text, the intuition behind "state" is not one of physical situations as in the light bulb example.

A HAL vector can be considered to represent the information "state" of a particular concept (or combination of concepts) with respect to a given corpus of text. The degree of information flow between "publisher" and the combination of "penguin " and "books" is directly related to the degree of inclusion between the respective information states represented by HAL vectors. Total inclusion leads to maximum information flow. Inclusion is a relation $\lhd$ over the vectors in the underlying HAL space.

Definition 2 (HAL-based information flow)

$i_1, \ldots, i_n \mid\!\sim j$ iff degree($\oplus c_i \lhd c_j$) $\geq \lambda$

where c_i denotes the conceptual representation of token *i*, and λ is a threshold value. (For ease of exposition, $\oplus c_i$ will be referred to as c_i because combinations of concepts are also concepts).

The degree of inclusion is necessary as HAL representations of concepts are exact representations, but are more akin to *characterizations* of the associated concept. Put simply, the representations are approximations and somewhat fuzzy. It is computed in terms of the ratio quality properties of c_i intersecting with properties in c_j to the number of quality properties in the source c_i:

$$\text{degree}(c_i \lhd c_j) = \frac{\sum_{p_l \in (QP_{\mu}(c_i) \cap QP(c_j))} w_{c_i p_l}}{\sum_{p_k \in QP_{\mu}(c_i)} w_{c_i p_k}}$$

The underlying idea of this definition is to make sure that a majority of the most important quality properties of c_i appear in c_j. When a threshold value 1.0 is set for λ, HAL-based information flow definition essentially equates to Barwise and Seligman's definition[6]. Note that information flow produces truly inferential character, i.e., concept *j* need not be a property dimension of $\oplus c_i$. (Examples of this will be given in section 5).

At the close of this section, we reflect back to Barwise and Seligman's definition of inferential information content given in the introduction and compare it to Definition 2. There is an obvious difference with respect to how the information inference problem is expressed syntactically. The syntax of Definition 2 has been carried over from Barwise and Seligman's Definition 1, whereby the information inference problem recast as one of information flow between tokens. This syntax circumvents using the rather clumsy syntax *r being F* etc. The syntax in terms of tokens is particularly useful for text based information inference as the tokens can be extracted from the text in question using standard techniques developed within the field of information retrieval. Concept combinations can be identified form text using part-of-speech tagging to identify noun phrases.

Definition 2 does not overtly make reference to the prior knowledge *k*. We argue that there is evidence to suggest that aspects of prior knowledge are captured by HAL. HAL is a co called global co-occurrence model. It acquires vector representations of meaning by capitalizing on large-scale co-occurrence information inherent in the text corpus. In other words, global co-occurrence models like HAL produce an *accumulated* vector representation of a word summed across the various contexts (windows) in which the word appears. Burgess and Lund state: "*We suspect that global co-occurrence models, more so than local co-occurrence models, will better capture the richness of cognitive and language effects that are important to the comprehension process.*" (Burgess & Lund 1997, p206). Moreover, there is evidence that the representations learned by HAL account for a wide variety of semantic phenomena (Burgess et al., 1998). On the basis of this, we consider the HAL space embody some aspects of *k*. Observe that *k* will change as the associated text corpus changes. A parallel can be drawn here with the Raynaud – fish oil discovery mentioned in the introduction. This discovery involves intermediate concepts connecting the two. When the bodies of disparate literature have reached a certain critical volume, sufficient *k* is available to allow the information inference between "Raynaud" concept and "fish oil" to proceed. In pragmatic terms, *k* is acquired by the process which constructs the HAL space. It

[6] The notion of inclusion is intuitively the same, but formalized differently: Barwise and Seligman define information flow in terms of a set of states being included within another set, whereas HAL-based information flow is defined in terms of how much one state (vector) is included in another.

is maintained by re-computing the HAL space as the associated text corpus changes.

Finally, Definition 2 can be placed within Gärdenfors' three level model of cognition. The left hand side of Definition 2 describes information inference between tokens at the symbolic level. These inferences are a result of computations within a cognitively motivated real-valued state space produced by HAL. We consider this space to be a primitive approximation of Gärdenfors' conceptual space. Moreover, due to the cognitive credentials of HAL, we put forward that the information inferences produced by Definition 2 may have a psychologistic character. Put more boldly, it is our hope that Definition 2 is an operational definition of a form of human reasoning with information, which is firmly in the mould of Barwise and Seligman's views of logic and cognition quoted at the end of section 2.

5. Information Inference in Practice: Inferring Query Models by Computing Information Flow

User queries to an information retrieval system are typically imprecise descriptions of the given information need. This phenomenon has been particularly emphasized with respect to queries on the web. Web queries average between two and three terms in length. Such short queries are, almost certainly, poor descriptions of the associated information need.

Various query expansion techniques have been developed in order to improve the initial query from the user. The goal of automatic query expansion is to automatically expand the user's initial query Q with terms related to the query terms in Q yielding a query Q'. The expanded query Q' is then used to return documents to the user. Various models and techniques have been proposed for determining the expansion terms. The language modelling approach to information retrieval has allowed query expansion to be considered as a language modelling problem. More specifically, a query language model comprises estimating the probability $P(t \mid Q)$ of every term t in the vocabulary in the light of the initial query $Q = (q_1, \ldots, q_m)$. Intuitively, those terms t with probabilities above a threshold can be considered more useful candidate terms for expanding the initial query Q. Two prominent query language modelling approaches are Relevance Model (Lavrenko & Croft, 2001) and Markov chain based model (Lafferty & Zhai, 2001). These will be used as reference points with respect to retrieval effectiveness.

Even though probabilistic approaches to query language modelling are promising, there are other points of departure. Query expansion can also be considered as an information inference process. If the terms $i_1, \ldots, i_m$ are query terms, then those terms j inferred informationally from these query terms can be considered as candidate query expansion terms. So, instead of a probabilistic foundation for the query language model via $P(t \mid Q)$, we propose a query language model based on the degree of information flow between the query Q and a vocabulary term t. The goals of the experiment reported below are twofold:

1. To ascertain the extent to which information inference improves retrieval effectiveness
2. To gain an understanding of the relative merits of using an inference based approach to query expansion versus probabilistic and semantic similarity based approaches

5.1 Experimental Set-up

The experiment used the AP89 collection (disk 1) together with TREC[7] topics 1 – 50. The collection contains 84,678 Associated Press documents. After removing stop words, a vocabulary of 137,728 terms resulted. Only the titles of the topics were used as queries (average query length: 3.24 terms).

Query processing

[7] TREC stands for the Text Retrieval Conference series run by NIST. See trec.nist.gov

In order to deploy the information flow model in an experimental setting, the query topics must analysed for concept combinations. In particular, the question of which concept dominates which other concept(s) needs to be resolved. As there seems to be no reliable theory to determine dominance, a heuristic approach is taken in which dominance is determined by multiplying the query term frequency (*qtf*) by the inverse document frequency (*idf*) value of the query term. More specifically, query terms can ranked according to *qtf*idf*. Assume such a ranking of query terms: $q_1, \ldots, q_{m.}$ ($m > 1$). Terms q_1 and q_2 can be combined using the concept combination heuristic described above resulting in the combined concept $q_1 \oplus q_2$, whereby q_1 dominates q_2 (as it is higher in the ranking). For this combined concept, its degree of dominance is the average of the respective *qtf*idf* scores of q_1 and q_2. The process recurses down the ranking resulting in the composed query "concept" $((..(q_1 \oplus q_2) \oplus q_3) \oplus \ldots) \oplus q_m)$. This denotes a single vector from which query models can be derived. If there is a single query term (m =1), it's corresponding normalized HAL vector is used for query model derivation. For each query topic, the query terms were combined in this way into a single query vector ($\ell_1 = 0.5, l_2 = 0.3, \alpha = 2.0, \partial_1 = 0, \partial_2 = 0$). This vector is then used to derive a particular query model .

As it is important to weight query terms highly, the weights of query terms which appeared in the initial query were boosted in the resulting query model by adding 1.0 to their score. Due to the way HAL vectors are constructed, it is possible that an initial query term will not be represented in the resulting query model. In such cases, the query term was added with a weight of 1.0. Pilot experiments show that the boosting heuristic performs better than the use of only query models without boosting query terms.

The query models

HAL spaces were constructed from the document collection using a window size of 8 words (l = 8). Stemming was not performed during HAL space construction, but stop words were ignored.

Composition Model (CM): In this model, the vector produced from the concept combination of the query terms is used as the query model. (The composition vector of *reagan⊕iran* given previously is an example of such a query model). The weights reflect the strength of association of the expansion term with the query term(s) averaged across all contexts in the collection. Essentially, the composition model represents a way of using only HAL vectors as a basis for query expansion. In addition, this model provides a reference point from which to gauge the effects of information inference.

Minkowski distance function (Mink): HAL produces a high dimensional space in which semantic distance can be computed. Gärdenfors states that the "similarity of two stimuli can be determined for the distances between the representations of the stimuli in the underlying psychological space" (Gärdenfors, 2000 p20). In practice, associations between words can be computed by calculating the similarity between their vector representations. The distance between two concepts *x* and *y* in the *n*-dimensional HAL space can be calculated using the Minkowski distance measure:

$$d(x,y) = \sqrt[r]{\sum_{1}^{n} (|x_i - y_i|)^r}$$

where $d(x,y)$ denotes the distance between the HAL vectors for *x* and *y*. Following Gärdenfors (2000), the similarity $s(x,y)$ between HAL vectors for *x* and *y* is calculated as an exponentially decreasing function of distance:

$$s(x, y) = e^{-c \cdot d(x,y)}$$

This model uses the query vectors to compute semantically related terms. More specifically, *x* represents a query vector and *y* represents the vector representation of an arbitrary term. If there is sufficient similarity, *y* is used as a query expansion term. In the experiment, the composition model (CM) is used to construct query vectors and the top 100 semantic associations and their associated weights are used to update the query vector, therby expanding it. The parameters were set as follows: *r*=2 and *c*=1/2350.The setting *r*=2 means the distance measured is a Euclidean distance, which is typical for information retrieval parameters. The *c* value affects the sensitivity of the similarity function. This value was determined empirically based on the collection being used.

Composition-based Information Flow Model (IM): This model computes information flows based on the query vectors using Definition 2. More specifically,

Given the query $Q = (q_{1,} \ldots, q_m)$, a query model can be derived from Q in the following way:

- Compute $\text{degree}(\oplus c_i \lhd c_t)$ for every term *t* in the vocabulary, where $\oplus c_i$ represents the conceptual combination of the HAL vectors of the individual query terms q_i, $1 \le i \le m$ and c_t represents the HAL vector for term *t*.
- The query model $Q' = \langle t_1 : f_1, \ldots, t_k : f_k \rangle$ comprises the top *k* information flows

Observe that the weight f_i associated with the term t_i in the query model is not probabilistically motivated, but denotes the degree to which we can infer t_i from *Q* in terms of underlying HAL space.

This model was chosen to investigate whether information flow analysis contributes positively to query model derivation. The top 85 information flows were used in the query model (*k*=85). This value produced best performance during a series of pilot studies.

Information Flow Model with pseudo-relevance feedback (IMwP): Pseudo-relevance feedback has consistently generated improved effectiveness. This model was implemented by constructing a HAL space by using the top fifty documents in response to a query, and thereafter deriving a query model from this local collection. The fifty documents were retrieved by the baseline model. The top 60 information flows were used in the query model (*k*=60). This value produced the best performance during a series of pilot studies.

Indexing , retrieval function and baseline model

Documents were indexed using the document term frequency and inverse collection frequency components of Okapi BM25 formula (Robertson et al., 1995) with parameters (k_1=1.2, k_2=0.0, *b*=0.75). Query vectors for the baseline model are produced using query term frequency with query length normalization (Zhai, 2001), which is defined similarly to the BM25's document term frequency with parameter $k_3 = 1000$. The matching function employed between document and query vectors was dot product as advocated by Lafferty and Zhai (2001).

Note that in the baseline model, terms were stemmed, whereas in the information flow models, terms were not stemmed, as pilot studies revealed that information flow models perform slightly better without stemming.

5.2 Results

This experiment evaluated the effectiveness of all models on the AP89 collection with TREC query topics 1-50. The results are as follows. The precision-recall curve for the Mink model was not depicted in Figure 1 as it is virtually the same as the composition model.

	Baseline	CM	Mink	IM	IMwP
AvgPr	0.182	0.197 (+8%)	0.193 (+6%)	0.247 (+35%)	0.258 (+42%)
InitPr	0.476	0.529 (+10%)	0.520 (+8%)	0.554 (+16%)	0.544 (+14%)
Recall	1687/3301	1996/3301 (+15%)	1655/3301 (-2%)	2269/3301 (+35%)	2331/3261 (+38%)

Table 2: Comparison of the query models against a baseline for the AP89 collection using TREC topics 1-50 (titles)

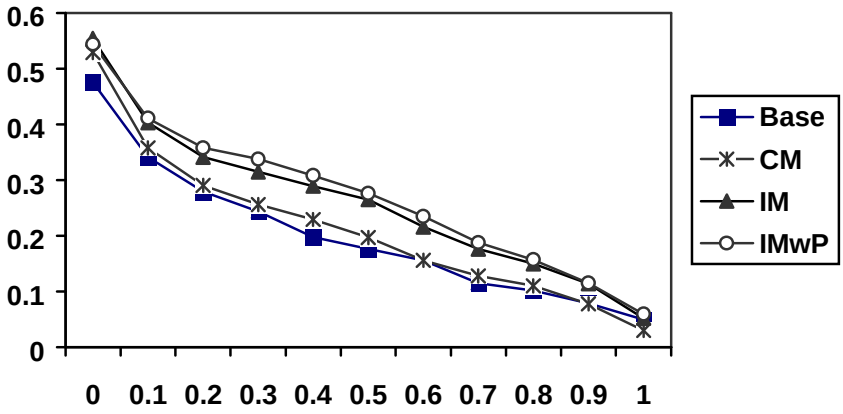


Figure 1: Precision recall curves comparing Baseline with the composition model, information flow model (with and without feedback) for the AP89 collection

5.3 Discussion

The first observation is the low baseline performance (average precision 0.182). This is due to only the titles being used as queries. The average precision score using the corresponding TDN (title, description, narrative) queries is 0.269.

The average precision of CM is 0.197. This represents an 8% improvement over the baseline. The IM model's average precision scored 0.247 which represents a 35% improvement over the baseline. Note that the difference between these models is the inference of expansion terms via information flow computation. As a consequence, we can conclude that the majority of the improvement with respect to precision was gained via information flow (27%). The improvement due to information flow is less pronounced with respect to recall: the CM model produced a 19% improvement (1996/3301 relevant documents retrieved) versus a 35% improvement in recall of the IM model (2269/3301). Therefore, the improvement due solely to information inference amounted to 16%. In summary, the inferential character of information flow contributes significantly to retrieval effectiveness, particularly with respect to precision. For example, part of the query model for "Space Program" (TREC topic 011) is as follows:

program:1.00 space:1.00 nasa:0.97 new:0.97 U.S.:0.96 agency:0.95 shuttle:0.95 national:0.95 soviet:0.95 president:0.94 bush:0.94 million:0.94 launch:0.93 called:0.93 **thursday:0.93** research:0.92 administration:0.92 flight:0.92 rocket:0.92 defense:0.91 **friday 0.91 project:0.91 system:0.91** mission:0.91 work:0.90 launched:0.90 officials:0.90 station:0.89 **long:0.88 announced:0.88** science:0.88 **scheduled:0.87 reagan:0.87** director:0.87 programs:0.87 air:0.87 **put:0.87** center:0.87 billion:0.87 aeronautics:0.87 **satellite:0.87 force:0.86 news:0.86 wednesday:0.86** technology:0.86 american:0.86 budget:0.86 **states:0.86 back:0.85 office:0.85 monday:0.85 plan:0.85 people:0.85** manned:0.85 **satellites:0.85 plans:0.84** development:0.84 **test:0.84 nation:0.84 mars:0.84 future:0.84** astronauts:0.84 **united:0.84 major:0.84 early:0.83 scientists:0.83 department:0.82** america:0.82 **laboratory:0.82 make:0.82 set:0.82 head:0.81** earth:0.81 house:0.81 **planned:0.81 tuesday:0.81 union:0.80 study:0.80 problems:0.80 april:0.80 earlier:0.80 ago:0.80 march:0.79 control:0.79 day:0.79 effort:0.79 money:0.79** star:0.79 **public:0.78 flights:0.78 develop:0.78 began:0.78 return:0.78 cost:0.78 pentagon:0.78 support:0.78 chief:0.77 moon:0.77 part:0.77 provide:0.77**

The bolded terms represent dimensions not present in the concept combination vector for *space⊕program.* In other words, they represent inferential information content with respect to the initial query "Space Program". Note that there is a number of inferences some of which appear to be related to the topic such as "satellite", "pentagon", "scientists" etc.

It is interesting to note that feedback in the information flow model produced a slight improvement of 4% average precision over the information flow model without feedback. This is a much smaller improvement than is typically the case when feedback is employed. Most of the improvement generated by the information flow model has taken place without feedback. This may due to the sparseness of the HAL space being generated from only fifty feedback documents. In other words, there is not sufficient text to produce "good" vector representations of terms. It may also be the case that the concept combination heuristic used on the global collection has contextualized the query vectors very effectively so there is less room for pseudo-feedback to generate improvements. More experimentation is needed to bear this out.

The semantic associations computed by the Minkowski distance metric did not produce any improvement over the composition model. This could mean that query expansion terms derived via computing information flow are more suitable than those computed via semantic similarity. More experiments will need to be conducted to bear this out, in particular, with respect to the Minkowski parameter (r = 1). This value has had some success in correlating vector similarity with cognitive effects (Burgess et al, 1998).

The experiment also allows comparison with the Markov chain query model which was investigated using the same experimental set-up described above (Lafferty & Zhai, 2001). With respect to average precision, the information flow model (IM) without feedback is 19% more effective than the Markov chain query model without feedback (0.247 vs. 0.201). The information flow model *without* feedback turns out to be slightly better (+6%) than the Markov chain model *with* feedback.

In other experiments (Bruza & Song, 2002), we compared the performance of HAL-based information flow with the Relevance model (Lavrenko & Croft, 2001). With respect to average precision, the information flow model scored higher average precision than the Relevance model for AP88&89 using topics 101-150 (0.301 vs. 0.261) and topics150-200 (0.344 vs. 0.319).

In summary, initial experimental results indicate that an information inference based approach to query expansion using HAL-based information flow generates encouraging improvements in retrieval effectiveness. These improvements are superior to two prominent probabilistic approaches using language models and a semantic similarity approach.

6. Summary and Conclusions

This article describes and evaluates an informational inference mechanism the theoretical basis of which is drawn from Barwise and Seligman's state-based information flow and Gärdenfors' cognitive model. The

theory is realized by computing information flow between vector-based representations of concepts derived from a text corpus by Hyperspace Analogue to Language (HAL). HAL vectors show compatibility with human information processing and are therefore interesting as computational representations of "meaning" which are cognitively, rather than logically motivated.

Since its inception, one of the goals of the logic-based information retrieval (IR) research agenda has been the development of a suitable model theory for IR. Initially, situation theory (Barwise, 1989) seemed to be a suitable candidate, but it has proven disappointing (see Wong et al. (2001) for an analysis). Even though situation theory is fundamentally a theory based on information, and inference can be defined in terms of channel theory (Van Rijsbergen & Lalmas, 1996), the potential of situation theory has never been realized. In fact, a recent theoretical analysis of situation theory has found it to have several implied negative characteristics (Wong et al., 2001). Part of the problem is situation theory being a symbolic theory with its attendant problems (e.g., computational complexity, difficulty in operationalizing its fundamental logical operators, etc). In addition, the representation of information (via infons) does not have any support from a cognitive perspective. The appeal of models such as HAL, latent semantic analysis (Landauer et al., 1998), and the like, is the evidence that the representations produced by these models accord with corresponding human representations. As a consequence, we feel that these form a more suitable basis for producing a model theory for IR and more generally as a promising basis for implementing systems which draw inferences which correlate with those a human would make.

The research presented in this article is motivated from a psychologistic perspective on information inference. The dimensional space derived by HAL can be considered to be a primitive approximation of the conceptual level of Gärdenfors' cognitive model. The input and output of the HAL-based information flow model are tokens, which are similar to keywords in IR. The inference mechanism, however, is driven via an equation computing vector inclusion within the state-space, thereby avoiding the computational complexities of inference at a symbolic level. The connection between the conceptual and symbolic level is furnished via a variation of Barwise and Seligman's definition of information flow. Two approaches were presented for priming vector representations according to context. The first approach uses a concept combination heuristic to adjust the vector representation of a concept in the light of the representation of another concept. The second approach computes a prototypical concept on the basis of exemplar trace texts. The prototypical concept vector can be contextualized by re-computing it with respect to a subset of traces which establish the context.

In a nutshell, we have presented the basis of a symbolic inference mechanism which is driven via computations between context-sensitive vectors at the conceptual level. We have therefore made a small step towards bringing Gärdenfors' cognitive model within the bounds of operational reality. Further research could be directed towards enhancing HAL to include more aspects of Gärdenfors' conceptual spaces, and to investigate information flow computations through other cognitively motivated vector representations of concepts.

HAL-based information flow was evaluated by measuring the effectiveness of query models produced by information inference using a corpus of news feeds to derive the conceptual space. Results show that HAL-based information flow contributes greatly to query model effectiveness, particularly with respect to precision. Moreover the results compare favourably with two prominent probabilistic models and another based on semantic associations computed via the Minkowski distance function.

More globally, this article details some preliminary groundwork for developing *semiotic-cognitive information systems (SCIS)* (Rieger, 1996; Newby, 2001; Gärdenfors & Williams, 2001;McArthur & Bruza, 2002). The term "semiotic-cognitive" refers to these systems manipulating "meanings" which are motivated from a cognitive perspective. Their ultimate goal is to enhance our cognitive firepower and thus help us become more aware in our ever more complex information environment.

Acknowledgments

The work reported in this paper has been funded in part by the Cooperative Research Centres Program through the Department of the Prime Minister and Cabinet of Australia.

References

Austin, J. (1998). Zen and the Brain: towards an understanding of meditation and consciousness. MIT Press.

Barwise, J. (1989). *The Situation in Logic.* CLSI Lecture Notes 17, Stanford, California.

Barwise, J., & Seligman, J. (1997) *Information Flow: The Logic of Distributed Systems.* Cambridge Tracts in Theoretical Computer Science 44.

Bruza, P.D., & Song, D. (2002). Inferring Query Models by Computing Information Flow. In Proceedings of the *ACM Conference on Information and Knowledge Management (CIKM).*

Burgess, C., & Lund, K. (1997) Parsing Constraints and High-Dimensional Semantic Space. *Language and Cognitive Processes, 12, pp.177-210.*

Burgess, C., Livesay, L., & Lund, K. (1998) Explorations in Context Space: Words, Sentences, Discourse, In Foltz, P.W. (Ed) Quantitative Approaches to Semantic Knowledge Representation. *Discourse Processes*, 25(2&3), 179-210.

Gärdenfors, P. (2000) *Conceptual Spaces: The Geometry of Thought.* MIT Press.

Gärdenfors, P., & Williams, M. (2001). Reasoning about Categories in Conceptual Spaces. In Proceedings of the *Fourteenth International Joint Conference of Artificial Intelligence*, pp. 385-392.

Gabbay, D. & Woods, J. Abduction. Lecture notes from the European summer School on Logic, Language and Information (ESSLLI 2000). Available: http://www.cs.bham.ac.uk/~esslli/notes/gabbay.html

Lafferty, J., & Zhai, C. (2001). Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In Proceedings of the *24th Annual International Conference on Research and Development in Information Retrieval (SIGIR'01), pp. 111-119.*

Landauer, T.K., Foltz, P.W., & Laham D. (1998) An Introduction to Latent Semantic Analysis. *Discourse Progress, 25(2&3), 259-284.*

Lavrenko, V., & Croft, W.B. (2001) Relevance-Based Language Models. In Proceedings of the *24th Annual International Conference on Research and Development in Information Retrieval (SIGIR'01), pp. 120-127.*

Lund, K., & Burgess, C. (1996) Producing High-dimensional Semantic Spaces from Lexical Co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28(2), pp. 203-208*

McArthur, R.M & Bruza, P.D. Mining tacit knowledge from small sets of utterance. Submitted. Available: http://www.dstc.edu.au/[to be filled in]

Newby, G.B. (2001) Cognitive Space and Information Space. *Journal of the American Society for Information Science and Technology, 52(12), pp. 1026-1048.*

Perfetti, C.A. (1998). The Limits of Co-Occurrence: Tools and Theories in Language Research. In Foltz, P.W. (Ed) Quantitative Approaches to Semantic Knowledge Representation. *Discourse Processes*, 25(2&3), 363-377.

Rieger, B. (1996) Situation Semantics and Computational Lingusitics, In Kornwachs, K. and Jacoby, K. (Eds). *Information: New Questions to a Multidisciplinary Concept.* Akademie Verlag.

Rijsbergen, C.J. van & Lalmas M., (1996) An Information Calculus for Information Retrieval. *Journal of the American Society for Information Science 47(5), pp 385-398.*

Robertson, S.E., Walker, S., Spark-Jones, K., Hancock-Beaulieu, M.M., & Gatford, M. (1996) OKAPI at TREC-3. In *Proceedings of the 3rd Text Retrieval Conference (TREC-3).*

Song, D., & Bruza, P.D. (2001) Discovering Information Flow using a High Dimensional Conceptual Space, In Proceedings of the *24th Annual International Conference on Research and Development in Information Retrieval (SIGIR'01), pp. 327-333.*

Swanson, D.R., & Smalheiser, N.R. (1997) An Interactive System for Finding Complementary Literatures: A Stimulus to Scientific Discovery. *Artificial Intelligence, 91, pp.183-203.*

Wong, K.F., Song, D., Bruza, P.D., & Cheng, C.H. (2001) Application of Aboutness to Functional Benchmarking in information retrieval. To be published by *ACM Transactions on Information Systems (TOIS),19(4), pp. 337-370.*

Zhai, C. (2001). Notes on the Lemur TFIDF model. Unpublished report.