AN INVESTIGATION INTO THE UTILITY OF THE PROBABILISTIC INTERPRETATION TO THE HYPER ANALOGUE TO LANGUAGE (HAL) SPACE.

LEIF AZZOPARDI

Abstract. In Song and Bruza [15], they introduce a framework for Information Retrieval(IR) based on Gardenfor's three tiered cognitive model; Conceptual Spaces[5]. They instantiate the conceptual space to generate informational inferences via the high dimensional conceptual space which are later used for ad-hoc retrieval. The implementation utilizes the Hyper Analogue to Language (HAL) algorithm [10] to build such a high dimensional conceptual space and generate inferences by employing Barwise and Seligman's theory of Information Flow[2].

In this report, we propose an alternative implementation of the conceptual space by using a probabilistic HAL (pHAL) space as the basis for the conceptual space. To evaluate whether converting to such an implementation is beneficial we have performed an initial investigation comparing HAL and pHAL using the concept combination process for query expansion. This was performed across a range of parameters that are involved in the construction of the (p)HAL space and combination techniques. The influence on the retrieval effectiveness when applying such methods to query expansion was the basis of comparison. Our results indicated that pHAL is a competitive alternative to the original HAL method.

Please note that this is a preliminary investigation which serves only to identify whether further analysis is warranted. The results reported in this study while not conclusive serve as an indication of the potential of a probabilistic alternative.

1. Introduction

A Conceptual Space is a model of cognition that views symbolic processing on three levels (see [5] for full details). The three tiers from higher to lower are Symbolic, Conceptual and Associationist. Each level represents cognition at a

Date: 1/9/2004.

different scale/resolution, and this varies greatly across levels. For instance an instantiation of the Associationist level is through connectionist systems such as neural networks while the Symbolic level could be represented through propositional logic. However, at the Conceptual level information is represented geometrically in terms of a high dimensional space and bridges the Symbolic and Associationist levels.

Within the domain of Information Retrieval(IR), Song and Bruza [15] offer an implementation of a Conceptual Space by employing the Hyper Analogue to Language (HAL) algorithm. HAL automatically constructs a high dimensional semantic space based on a corpus of text which appears ideal for the implementation of the conceptual space. HAL as presented by Song and Bruza [15] is broken into several processes:

(1) generation of the high dimensional semantic space using HAL;
(2) from this space concepts are combined to form high level concepts;
(3) the quality properties (those relevant to the context) given this higher level concept are selected, resulting in the final *combined concept* and;
(4) this combined concept is then used to draw inferences about other concepts (based on geometric functions) generating *information flows*.

Either the combined concepts or information flows are then used to expand the query.

We propose a probabilistic interpretation of the Hyper Analogue to Language as it provides a principled framework in which the inferences drawn from the conceptual space can be seamlessly integrated into a retrieval model. Also, it provides the possibility of employing a range of techniques developed in other areas such as Statistical Natural Language Processing that would otherwise be unavailable.

In this report, we provide a preliminary comparison between the original HAL method and our proposed probabilistic HAL. The remainder of the report is as follows: We describe each process involved in the construction of the conceptual space and a probabilistic variant. Then we examine a range of different parameters which affect the conceptual space. The differences are evaluated with respect to the IR task of query expansion. From these preliminary results we discuss our findings and their implications, before detailing the limitations and the further work.

2. Defining the Conceptual Space

2.1. **Hyper Analogue to Language.** The Hyper Analogue to Language (HAL) Algorithm [10] was developed as a representational model of semantic memory and has been adapted for the purposes of Information Retrieval [15]. The intuition underlying HAL is that when a human encounters a new concept they derive its meaning from accumulated experience in the context in which the concept appears. Thus, the meaning of a concept can be learnt from its usage with other concepts within the same context. The representational model of this process is constructed automatically from a high dimensional semantic space over a corpus of text [4, 13]. For the construction of the space individual concepts (terms) are used as the basis. These concepts can be combined to form new higher order concepts.

The construction of the HAL space [10] can be described as follows: each term t in the vocabulary T is composed of a high dimensional vector over the T, resulting in a $|T|$ by $|T|$ HAL matrix, where $|T|$ is the number of terms in the vocabulary. A window of length K is moved across the corpus of text at one term increments ignoring punctation, sentence and paragraph boundaries. All terms within this window are said to co-occur with the first term in the window with strengths inversely proportional to the distance between them. The weighting assigned to each co-occurrence of terms is accumulated over the entire corpus. The weighting scheme and window size is based on the cognitive limitations of humans.

Mathematically, we can formalize this as: for a term t and any other term t' the HAL co-occurrence score can be represented as the weighted sum over all the different window lengths. Where $n(t,k,t')$ is the number of times term t' occurs k distance away from t, and $w(k) = K - k + 1$ denotes the strength of relationship between the two terms.

$$HAL(t'|t) = \sum_k w(k)n(t,k,t') \tag{2.1}$$

The length of the window size will invariably influence the quality of the associations within each HAL vector. For instance, as the size of the window increases, the higher the chance of representing spurious associations between terms. Various windows sizes have been used from 2 to 10. However, it is unclear what is the best size of window given the purposes of IR, though Song and Bruza[15] suggest a bi-directional window of 8. This is a modification of the original HAL Space which is direction sensitive because it records the co-occurrence information for terms preceding every term. It was found that preserving this term order was not useful for IR and the combination of the row and column vectors for a term (thus a bi-directional window) was more effective [15]. For instance with the sentences "The black cat ..." and "The cat is black.", while the ordering is different the notion that the cat is a particular colour, black, is preserved when taking both directions into account. However, this may not always be appropriate and could be detrimental to the representation.

2.2. **Probabilistic HAL.** Our proposed implementation of a probabilistic HAL space is not entirely new (see [7, 9]), however our formalization is slightly different to those already proposed. The HAL Space naturally lends itself to a probabilistic interpretation as term co-occurrence counts can be used to define conditional probabilities. This can be interpreted as given the term t, what is the probability of associating term t' with term t given the window of size k? The association is defined by the weighting function $w(k)$ which can be replaced with a prior denoting the strength of the co-occurrence based on the distance within the window K.

$$pHAL(t'|t) = \sum_k p(k)p(t'|t,k) \tag{2.2}$$

where $p(t'|t,k) = \frac{n(t,k,t')}{\sum_{t'} n(t,k,t')}$.

To ensure that the we obtain a valid probability distribution, the constraint that $\sum_{t'} p(t'|t) = 1$ is imposed. A similar formation of the conditional probability is suggested in [7], though without the prior $p(k)$. Such a prior can now be assigned different weightings or estimated.

3. Concept Combination

3.1. **using HAL Spaces.** Once a HAL Space is constructed for a corpus of text, Song and Bruza [15] apply an ad hoc method for combining the concepts formed in the HAL Space. The combination of concepts produces a higher order concept that is represented as another concept HAL vectors.

Given a query Q comprising of query terms $q_1, \ldots, q_l$, the process of concept combination is performed to build a new concept vector effectively expanding the query based on the co-occurrence information amassed in the HAL Space. In order to do so, the query terms are ordered by the amount of information the term carries according to the query term frequency by Inverse Document Frequency Score (i.e QF.IDF). This establishes the dominance of each query term. Thus the ordering is from the most specific term to the most general. There is no formal justification for using such a weighting and there may be more appropriate ways of ordering the set of query terms that capitalizes on the structural relationship between terms. For instance grammar trees, Maximum Spanning Trees or the natural order are all possibilities.

The following steps are applied to combine the ordered set of query terms $q_1, \ldots, q_l$:

3.1.1. *Step 1.* The concept vectors defined by the HAL space given the query terms are re-scaled as weighted concept vectors, $wc(t|q_i)$.

$$wc_1(t|q_1 \oplus ... \oplus q_i) = w_1 + \frac{w_1 * HAL(t|q_i)}{max_{t'} HAL(t'|q_i)}$$

$$wc_2(t|q_{i+1}) = w_2 + \frac{w_2 * HAL(t|q_{i+1})}{max_{t'} HAL(v'|q_{i+1})}$$

where the weights $w_1, w_2 \in (0...1)$ and $w_1 > w_2$. It is suggested that these weights are set to $w_1 = 0.6$ and $w_2 = 0.4$ from empirical findings[3].

3.1.2. *Step 2.* The weights of terms v appearing in both concepts wc_i and wc_{i+1} are strengthen via a multiplier α, where $\alpha > 1$.

$$\forall (t \in (wc_1 > 0) \wedge t \in (wc_2 > 0)|$$

$$wc_1(t|q_1 \oplus ... \oplus q_i) \leftarrow \alpha * wc_1(t|q_1 \oplus ... \oplus q_i)$$

$$wc_2(t|q_{i+1}) \leftarrow \alpha * wc_2(t|q_{i+1})$$

3.1.3. *Step 3.* Combined the weighted concept vectors $wc_1 \oplus wc_2$:

$$wc_1(t|q_1 \oplus ... \oplus q_{i+}) = wc_1 + wc_2$$

3.1.4. *Step 4.* The new weighted concept vector is normalized to unit norm. And steps 1 to 4 are repeated for each addition query term.

$$wc_1(t|q_1 \oplus ... \oplus q_{i+1}) = \frac{wc_1(t|q_1 \oplus ... \oplus q_{i+1})}{\sqrt{\sum_{t'} wc_1(t|q_1 \oplus ... \oplus q_{i+1})^2}}$$

If the number of query terms is one, then the HAL vector $HAL(t|q_1)$ is used to describe the concept.

3.1.5. *Step 5.* Boost the query terms; add γ to the $\forall q_i \in Q | wc_1(q_i|q_1 \oplus ... \oplus q_l) \leftarrow wc_1(q_i|q_1, ..., q_l) + \gamma$. The suggested value for γ is 2, again set from empirical findings [3].

Fundamentally, the process attempts to create a representation of the context in which the query terms are used and attempts to remove ambiguous contexts through the re-weighting/strengthening of common terms in the concept vector. The process is somewhat heuristical based but intuitively grounded.

3.2. **using pHAL.** From our pHAL space we can define a methods of combination in a principled manner that stems directly from the pHAL space. Whilst some of the intuition behind the combination is lost, the proposed methods attempt to mimic the original concept combination process in a simple fashion as a starting point.

3.2.1. *Method 1.* Our first proposed method is a mixture model of the query and their pHAL representations to generate a query model θ_Q. The dominance of each query term is encoded using the prior λ_i. The query model $p(t|\theta_Q)$ can be defined as follows:

$$p(t|\theta_Q) = \lambda_0 p(t|Q) + \sum_1^l \lambda_i pHAL(t|q_i) \tag{3.1}$$

where $p(t|Q)$ is the empirical probability of term t given the query Q, $p(t|q_i)$ is the probability of a term t given the query term q_i as defined by the pHAL Space $pHAL(t|q_i)$, and the constraints $\sum_0^l \lambda_i = 1$ where $\lambda_0 > \lambda_1 > \cdots > \lambda_i > \cdots > \lambda_l$ are imposed.

Notice that this interpretation differs from the original as there is no attempt to disambiguate the context in which a term occurs such as is done in step 2 of the original method. For instance, given the concept 'reagan', as in President of the

United States, Ronald Reagan, the context in which this concept appears are varied including domestic affairs, Japanese trade disputes, Middle East controversy, etc. Each context contributing to the terms that appear in the co-occurrence representation for 'reagan'. The problem is to select the appropriate context. The original method takes a positive approach to this problem by re-weighting the terms that are shared between the concepts. This lowers the weight of those terms that are not in the same context.

3.2.2. *Method 2.* We employ a different approach adapted from Amati [1] where the context terms are filtered. To select terms to expand a query, Amati employed a simple boolean condition to select terms from the top n retrieved documents. The selection criteria was that if at least two of the n documents contained the term then it is kept, otherwise it is discarded. We adapt the condition and apply it as follows: A term is kept if it appears in at least two of the l pHAL representation for the query terms. The updated pHAL representations for each query term is re-normalized to ensure that it is a probability and the combination is performed as in Method 1.

3.2.3. *Other Alternatives.* Potentially, there are numerous ways to combine the information provided given the query terms. Incorporating some decision mechanism to select the appropriate context terms to use in the combination would be more intuitive for building an expanded representation of the query, perhaps using Jeffery's rule of conditionalization, a decision theoretic framework, parsimonious language models, etc. Alternatively, adapting query expansion as proposed in Amati's framework could potentially build better combinations of concepts. Or further, encoding the structural dependencies between terms (using Maximum Spanning Tree, subsumption, etc) could produce other variants.

4. Selection of Quality Properties

Once the concept combination is performed selection of quality properties is then applied - to choose those terms in the representation that are the most important.

4.1. **Mean Thresholding.** Quality properties are defined as those properties $HAL(v|Q) > \delta$, where δ represents a threshold which is usually set to the mean of the vector $HAL(.|Q)$. As previously mentioned properties which are under the mean weight of the vector are removed. This technique removes those co-occurrences which appear very infrequently. However, it does not remove terms which are very common and from an Information Retrieval point of view may not be particularly useful in discriminating between relevant and non-relevant documents.

4.2. **Log-Likelihood Ratio.** - Employing the Log-Likelihood ratio (LLR) provides a principled method for selection of quality properties. This selection criteria has been used in the statistical Natural Language Processing to detect collocations[11]. It was also proposed as part of a theory of semantic space[8], where relationships between terms within a window of text could be identified (akin to the (p)HAL space). This was performed by using the odds ratio to determine whether term t_i is dependent on t_j versus term t_i occurring independently of t_j. This can be expressed as two hypotheses (Dunning,1993):

- Hypothesis 1. $p(t_i|t_j) = p_1 \neq p_2 = p(t_i|\bar{n_j})$
- Hypothesis 2. $p(t_i|t_j) = p = p(t_i|\bar{t_j})$

Hypothesis 1 is a formalization of dependence which is good evidence that the two terms are indeed related. Hypothesis 2 is the formalization of two terms occurring independently. The log odds ratio $\log \frac{H_1}{H_2}$ gives the odds that the terms are dependent if greater than zero and independent if less than zero. We selected the LLR as opposed to other methods because it is robust to small frequencies and gives a clear selection criteria. Though there are numerous other methods that can be employed (see Section 4.4).

4.3. **Examples of Selections Techniques.** To show the difference between the two techniques we provide several examples of the application of each technique (See Table 1). Given a term t and its corresponding pHAL vector $pHAL(t'|t)$, selection is performed with both techniques. The terms which meet the selection criteria are kept the rest are discarded, the vector is then re-normalized to ensure that the pHAL vector is a probability distribution. The following examples have been generated from a pHAL Space with window size k=5 and uniform weighting using the Wall Street Journal Collection (see Section 5 for further details). The terms in the examples given are word stems. The top twenty terms are displayed in decreasing order of probability for each selection method. Note that since the Mean

Mean manufactur	LLR manufactur	Mean reagan	LLR reagan	Mean expense	LLR expense	Mean dirt	LLR dirt
hanov	hanov	administr	administr	said	past	road	road
said	product	mr	bush	million	30	mr	track
product	manufactur	presid	bill	compani	dai	track	cheap
compani	plant	year	offici	past	cost	sai	poor
corp	trust	said	veto	30	relat	new	long
new	good	bush	polici	dai	incom	cheap	formula
industri	sector	bill	sign	year	account	poor	20
oper	equip	offici	hous	mr	deduct	long	bag
manufactur	retail	veto	budget	cost	ratio	formula	dirt
year	order	polici	aid	new	expens	20	find
mr	associ	sign	reagan	sai	incur	said	bike
market	job	sai	white	quarter	currenc	bag	pile
1	engin	hous	congress	1	paid	us	floor
plant	capac	budget	democrat	oper	model	dirt	south
unit	facil	aid	defens	increas	fee	find	409
trust	export	new	campaign	relat	item	bike	hour
sai	process	reagan	econom	incom	percentag	pile	marbl
concern	factori	white	soviet	sale	provis	floor	hand
good	output	tax	cut	2	hurt	take	gold
2	distribut	congress	polit	account	litig	south	don
million	chemic	democrat	appointe	fund	taxpay	two	help

Table 1. Top twenty terms for the terms 'manufactur', 'reagan', 'expense' and 'dirt'.

Thresholding method removes terms that have a frequency less than the mean, the high frequency terms remain. Under the Log-likelihood method each term is selected on the basis that it is dependent on the initial term. As a result a more intuitive vector results, which appears to be much *cleaner* than the former method. By cleaner, we mean that words that you would not intuitively associate with the pivot term have been removed even though they appear at high frequency with the pivot term. For instance with all the terms, associations such as mr, say, said do not automatically spring to mind (of course this is rather subjective). These terms are rather uninformative (according to their Inverse Document Frequency (IDF) Score) as they appear in most news articles in the collection. Perhaps a case may be built to remove such common words (and treat them like stop words). However there are other terms which are removed from the vector, only six of the top twenty terms given 'manufactur' have been kept by the LLR method. So while the other terms occur frequently with 'manufactur' the odds ratio suggests that the dependency between the two terms was not likely because such terms are likely to be produced independently of seeing the term 'manufactur'. Hence there is no significant dependency. This has removed terms that we may associate with 'manufactur' such as 'compani', 'corp' and 'market', the effect this has on IR performance is unknown, but as these terms are relatively non-informative (according to their IDF score) their ability to discriminate relevant from non-relevant is probably quite low.

It appears that the Log-likelihood Ratio selection criterion produces pHAL vectors which intuitively seem more appropriate as the associations that are very frequent and occur as often randomly in the corpus are removed.

4.4. **Alternative Methods.** An array of methods can be performed on the probabilistic HAL space (see [9, 11]) for the selection of such quality properties. For instance, t-test, mean and variance, chi square test and pointwise mutual information measures.

5. Empirical Study on the Wall Street Journal

The focus of this study was to examine whether a probabilistic implementation could perform as well or better than the original method proposed. In our analysis we have also considers the different factors that may affect the quality of the resulting p(HAL) spaces and its effect on the retrieval process. We have attempted to assess these factors with respect to the task of query expansion and measured the influence on IR effectiveness in terms of mean Average Precision (mAP). We also considered the precision at ten percent, however the results were similar to those obtained for mAP. For brevity we have excluded these. The experimental methodology was as follows:

We took approximately 40000 documents from the Wall Street Journal Collection. These were indexed, standard stop words were removed and Porter Stemming[14] applied. The size of the vocabulary was $|T| = 30239$. For each term in the collection the number of times term t_i occurred with t_j at a distance of length k, $n(t_i, k, t_j)$, was recorded. This data was then processed to compute the various (p)HAL structures.

5.1. **Generating (p)HAL Spaces.** Specifically, we examined the following variables:

- k - the size of the window size; as the window size k is increased the greater the chance of capturing spurious relationship and this is known to influence the IR performance. Due to limitations in memory we only tried the window sizes of one, three and five with a bi-directional window.
- $p(k)$ - the weighting assigned to the term co-occurrence. We examined two possibilities.
  - Linear - the weighting assigned between two terms $p(k)$ is proportional to the distance between them.
  - Uniform - the weighting assigned is equal regardless of the distance between terms.
- Selection of Quality Properties - Two methods for selecting quality properties were employed; Mean Thresholding and Log-Likelihood Ratio. Whilst in the original method selection is performed after combination, we have applied selection at before the concept combination process and then performed ad-hoc querying and then after concept combination and then performed ad-hoc querying. The rationale for performing a before and after analysis is to determine the influence the selection process is having on the overall query expansion / retrieval process.

5.2. **HAL Concept Combination.** The following parameter values were set for combining concepts: $w_1 = 0.6$, $w_2 = 0.4$, $\alpha = 2$ and $\gamma = 2$. For the purposes of our experiments, the resulting concept combination from the HAL space $HAL(t|Q)$ was normalized to sum to one. This defined the probability distribution for the probability of a term given the HAL query model $p_{HAL}(t|\theta_Q)$. This was used to when selecting the quality properties using the Log-Likelihood selection method. In this investigation we did not consider the Information Flow component which attempts to find semantically related terms given the context defined by the $p_{HAL}(t|\theta_Q)$. This is left for further work.

5.3. **Query Expansion with Concept Combination.** Given the constructed (p)HAL Space we performed query expansion given the initial query terms. We used the titles of the TREC Topics 101-150 as initial queries and used the concept combination methods to build a query model θ_Q. The top n terms from the query model were then used as an expanded query. We used a standard Language Modelling approach[16] where we attempt to predict the query Q given the document model θ_d.

$$p(Q|\theta_d) = \prod_{t \in Q} p(t|\theta_d)^{n(t,Q)} \tag{5.1}$$

In the cases where we expanded the initial query Q, we used the query model θ_Q and replace $n(t,Q)$ with the query model estimate $p(t|\theta_Q)$ computing the score over the top n terms in the query model.

Document models were constructed with Bayes Smoothing [16] (see Equation 5.2) where $n(t,d)$ is the number of times term t occurs in document d, $n(d) = \sum_t n(t,d)$ is the total number of times in d, $p(t) = \frac{\sum_d n(t,d)}{\sum_{d'} n(d')}$ is the probability of the term given the collection and β is the free parameter.

$$p(t|\theta_d) = \frac{n(t,d) + \beta p(t)}{n(d) + \beta} \tag{5.2}$$

6. Results

The experimental factors that we have manipulated in the course of this experiment are as follows: Space Window Size (k=1,3,5), Space Weighting (w= Uniform, Linear), Number of query expansion terms (et= 5,10,50,85,100), Concept Combinations (c = HAL, pHAL-1, pHAL-2), Bayes Smoothing Parameters (β = 1000,2500, 5000) and selection (s) (Before (mean, Log Likelihood Ratio) and After (mean, Log Likelihood Ratio)). Whilst we have been rather restrictive over the range of parameters chosen even this means that the number of possible conditions given a query is in the order of 10^3. To aid analysis we have partitioned these factors into two groups; representational factors that influence the quality of the query model generated (though this is quantified in terms of retrieval performance) and retrieval factors that that are used and required by the retrieval model.

The representational factors are: window size(k), space weighting (w) and the selection mechanism (s). The retrieval factors are: the query model constructed by the different concept combination methods (c), the number of expansion terms (et), and the smoothing parameter (β).

6.1. **Window Size.** The window size used in the construction of the (p)HAL Space has an noticeable influence on the overall IR performance (see Table 2). As the window size increases, an increase in IR performance is witnessed; this is regardless of the concept combination method employed or when quality property selection is performed. The IR performance increase appears to be tapering off with the increase in window size; this can be quantified by examining the mean difference between k=3 and k=1, versus k=5 and k=3, where the mean difference is 1.68 percent and 0.42 percent, respectively. Though further analysis at higher k is required to determine the optimal size as other studies have shown higher window sizes produce better results.

6.2. **Space Weighting.** There appears to be a marginal difference in IR performance given the two different weighting schemes (see Table 3). Surprisingly, the linear weighting scheme employed in the construction of the original HAL in [10] performed slightly worse than simply employing a uniform weighting scheme.

Given the results from the window size (k) and weighting function (w), we confirmed that the best performance was obtained when the constructed (p)HAL Space with k=5 and an uniform weighting (see Table 4).

6.3. **Selection of Quality Properties.** Depending on the selection technique used and when, this will impact on the quality of the concept combination. Performing selection before the application of the concept combination reduces the amount of term co-occurrence information available (and this saves a lot of storage space - see Table 5). By reducing the amount of information available by selecting the quality properties for each (p)HAL vector before hand we hoped that the remaining information was of higher quality and would enable better concept combinations. From our results however this was not the case. Applying either method

Sel.	Comb.	k=1	k=3	k=5
Mean	HAL	23.6	24.7	**25.1**
Before	pHAL-1	23.2	25.2	**25.6**
	pHAL-2	22.0	25.0	**25.7**
Mean	HAL	24.4	25.1	**25.4**
After	pHAL-1	24.2	25.6	**25.7**
	pHAL-2	23.6	25.4	**25.8**
Log	HAL	24.3	25.5	**25.8**
Likelihood	pHAL-1	23.7	25.3	**25.6**
Before	pHAL-2	20.1	22.9	**23.8**
Log	HAL	24.9	25.8	**26.1**
Likelihood	pHAL-1	24.4	25.9	**26.2**
After	pHAL-2	23.7	25.8	**26.3**

TABLE 2. The mean Average Precision of TREC Titles 100-150, where the mean is taken over all queries, weighting functions (w), expansion terms (et) and smoothing parameters (β) for each window size (k)

before combination resulted in slightly poorer performance though the difference was marginal. On average the LLR method performed after combination obtained the best IR performance (see Tables 2 , 3 and 4).

So far in our analysis we have examine the main over arching factors that influence the IR effectiveness. This has shown that the best results obtained tend to be when k=5, the weighting is uniform and selection of the quality properties is done after concept combination. For the subsequent sections, we shall restrict the set of results for analysis to this subset as this set consistently delivers the best retrieval performance across all the concept combination methods (HAL, pHAL-1 and pHAL-2). Where appropriate we also report HAL-O, which is the HAL Space implemented according the original method.

6.4. **Expansion Terms.** From Table 6, as the number of query terms is increased the IR Performance steadily increases. The best results obtained tend to be around 85 to 100 additional terms. We performed a few extra runs at higher numbers

Sel.	Comb.	Uniform	Linear
Mean	HAL	**25.0**	24.8
Before	pHAL-1	**25.5**	25.2
	pHAL-2	**25.5**	25.2
Mean	HAL	**25.3**	25.2
After	pHAL-1	**25.8**	25.6
	pHAL-2	**25.8**	25.4
Log	HAL	**25.7**	25.3
Likelihood	pHAL-1	**25.7**	25.2
Before	pHAL-2	**23.8**	22.9
Log	HAL	**26.1**	25.9
Likelihood	pHAL-1	**26.2**	25.9
After	pHAL-2	**26.3**	25.9

TABLE 3. The mean Average Precision of TREC Titles from Topics 100-150, where the mean is taken over all queries, window size (k), expansion terms (et) and smoothing parameters (β) for each weighting scheme (w) employed

of additional terms however the mAP appeared to plateau after about 100 terms. This is not surprising given that the probability mass assigned to subsequent terms becomes progressively smaller, hence the effect on the overall document ranking is lessened.

6.5. **Comparison.** Query Expansion has been extensively studied in the IR literature. The general conclusions from this body of work is that automatic query expansion is often highly effective for many information retrieval tasks when a short query is submitted by the user. However, there are cases when retrieval performance may be degraded when query expansion techniques are applied, for instance when early precision is critical or the number of relevant documents is small. Query expansion usually results in a gain in recall but this is accompanied by a loss in precision[6]. Approximately one-third of expanded queries will suffer a drop in average precision [12]. While an increase in mean Average Precision can be obtained

Sel.	Comb.	Weight	k=1	k=3	k=5
Mean	HAL	Uniform	23.6	24.8	**25.2**
Before	HAL	Linear	23.6	24.6	25.0
Mean	pHAL-1	Uniform	23.2	25.3	**25.7**
Before	pHAL-1	Linear	23.2	25.0	25.4
Mean	pHAL-2	Uniform	22.0	25.2	**25.9**
Before	pHAL-2	Linear	22.0	24.8	25.6
Mean	HAL	Uniform	24.4	25.2	**25.5**
After	HAL	Linear	24.4	25.1	25.3
Mean	pHAL-1	Uniform	24.2	25.2	**25.5**
After	pHAL-1	Linear	24.2	25.1	25.3
Mean	pHAL-2	Uniform	23.6	25.6	**25.9**
After	pHAL-2	Linear	23.6	25.2	25.7
LLR	HAL	Uniform	24.3	25.6	**25.9**
Before	HAL	Linear	24.3	25.3	25.8
LLR	pHAL-1	Uniform	23.7	25.5	**25.9**
Before	pHAL-1	Linear	23.7	25.1	25.4
LLR	pHAL-2	Uniform	20.1	23.4	**24.2**
Before	pHAL-2	Linear	20.1	22.5	23.3
LLR	HAL	Uniform	24.9	25.9	**26.2**
After	HAL	Linear	24.9	25.7	26.0
LLR	pHAL-1	Uniform	24.4	26.0	**26.3**
After	pHAL-1	Linear	24.4	25.7	26.1
LLR	pHAL-2	Uniform	23.7	26.0	**26.5**
After	pHAL-2	Linear	23.7	25.6	26.2

Table 4. The mean Average Precision of TREC Titles from Topics 100-150, where the mean is taken over all queries, expansion terms (et), smoothing parameters (β) for each weighting scheme (w) and window sizes (k) employed. Bold results show the best mAP given w and k.

Sel.	k=1	k=3	k=5
After	3227326 (100)	8481697 (100)	12417111 (100)
Before - Mean	1425257 (44.2)	3844673 (45.3)	5648082 (45.7)
Before -LLR	2018041 (62.3)	4412774 (52.0)	6137006 (49.4)

Table 5. Number of values stored in the HAL matrix before and after selection. The figure in brackets is the percentage of the original.

Sel.	Comb.	et=5	et=10	et=50	et=85	et=100	mean et=50
Mean	HAL	24.8	24.9	25.8	26.0	25.9	25.5
After	pHAL-1	24.8	24.9	25.8	26.0	25.9	25.8
	pHAL-2	24.9	25.3	26.4	26.7	26.4	25.9
Log	HAL	24.9	25.3	27.1	26.8	27.0	26.2
Likelihood	pHAL-1	25.2	25.7	26.8	26.9	27.0	26.3
After	pHAL-2	25.3	25.8	27.0	27.1	27.3	26.5

Table 6. Mean Average Precision given the extra terms used in the query.

this is not usually uniform across all queries[1]. With these conclusions in mind, we have performed the following analysis.

We have selected for comparison the (p)HAL spaces where k=5 with uniform weighting, using the Log Likelihood Ratio to select quality properties after combination. For the query expansion, the number of extra terms for HAL, pHAL-1 and pHAL-2 was set to 100. The HAL configuration k=5 with linear weighting with mean thresholding applied to select quality properties is also shown (HAL-O). We assumed that the best smoothing parameter can be selected for each of the models (standard or otherwise, across the range tested). In addition to comparing the different expansion models (HAL, pHAL-1, pHAL-2) against the baseline method with just the short queries, we have included as a human query expansion (HQE) the results using the baseline method and the entire TREC Topic (See Figure 1).

Aside: Ideally, from our conceptual space we hope to derive human like information inferences to expand the query (or generate a query model). Whilst we have

<top>
<head> Tipster Topic Description
<num> Number: 105
<dom> Domain: Finance
<title> Topic: 'Black Monday'
<desc> Description:
Document will state reasons why U.S. stock markets crashed on 19 October 1987 ('Black Monday'), or report on attempts to guard against another such crash.
<smry> Summary:
Document will state reasons why U.S. stock markets crashed on 19 October 1987 ('Black Monday'), or report on attempts to guard against another such crash.
<narr> Narrative:
A relevant document will contain at least one reason why U.S. stock markets experienced a huge price drop on 19 October 1987, losses of equity so large that markets were said to have crashed (the Dow, for example, lost 508 points on that one day alone); the date of the crash has become known as "Black Monday." A preferable document would contain a detailed analysis of the crash. The best document would link analysis of events to actions taken or recommendations made by federal authorities or the stock markets to prevent future crashes. NOT relevant are reports which simply reference, without analysis, "Black Monday," such as anniversary stories generated by the press around every October 19th.
<con> Concept(s):

(1) 19 October 1987, 'Black Monday'
(2) New York Stock Exchange, Securities and Exchange Commission, SEC, National Association of Securities Dealers
(3) Chicago Board of Trade, Chicago Mercantile Exchange, Commodity Futures Trading Commission, CFTC
(4) program trading, index arbitrage, futures market, portfolio insurance, specialists, margins, super DOT system
(5) Brady Commission, circuit breaker

<fac> Factor(s):
<nat> Nationality: U.S.
<time> Time: any time after 19 October 1987
< /fac>
<def> Definition(s):
< /top>

FIGURE 1. An example of a TREC Topic 105, in our experiments we used the Title as the query, to simulate a human expanded query (HQE) we used the text in the entire TREC topic from the fields, Title, Description, Summary, and Concepts.

not employed the informational inference component in this study, the performance gained (if any) from a better description of the information need provides a test to determine if the model can generate human like expansions[1]. Also, this invites an interesting opportunity to examine whether the human expansion will perform in a

[1]This would be the IR equivalent of pitting a human against a computer at chess, for the task of query expansion

	Baseline	HAL-O	HAL	pHAL-1	pHAL-2	HQE
TOPICS 101-150	25.3	26.3	27.4	27.0	27.9	40.9

Table 7. Mean Average Precision

	WIN	LOSE	DRAW
HAL-O	29 (58)	15 (30)	6 (12)
HAL	29 (58)	17 (34)	4 (8)
pHAL-1	31 (62)	14 (28)	5 (10)
pHAL-2	30 (60)	14 (28)	6 (12)
HQE	39 (78)	8 (16)	3 (6)

Table 8. Number of wins, loses and draws over the baseline method. The figure in brackets is the percentage.

similar manner to machine expansion (i.e failing one third of the time, non-uniform gain in AP, etc.).

We performed a query-wise Wilcoxon rank sum test ($\alpha = 0.05$) to determine whether a significant improvement in performance was achieved. All the expansion methods significantly outperformed the baseline of the short queries. The expansion methods, HAL-O, HAL, pHAL-1 and pHAL-2, were not significantly different from each other. However, the HQE was significantly better than all other methods. Thus, the overall ranking of techniques: Baseline $<$ HAL-O $\approx$ HAL $\approx$ pHAL-1 $\approx$ pHAL-2 $<$ HQE.

In Table 8, we show the number of wins, losses and draws for each method versus the baseline method. The (p)HAL based expansion techniques fails to exceed the baseline approximately a third of the time which is consistent with previous literature on query expansion techniques[12]. However, the human query expansion has performed somewhat better, failing only approximately one sixth of the time. This also shows that a non-uniform increase in mean Average Precision has resulted with the usage of these expansion techniques.

7. Discussion and Conclusion

Throughout this analysis we have examined a host of parameter values and settings with in the (p)HAL space that affect the IR effectiveness with respect to the task of query expansion. Our results indicate that our alternative implementation of the HAL space is comparable with the original method under the task assessed.

We have examined different weighting schemes and found that a uniform weighting scheme provided better results on average than the cognitively motivated linear weighting scheme, though these results were not statistically significant. We speculate that this is because as the window size increases the weighting assigned to distant terms is significantly smaller. i.e when using a bidirectional window of size five, the difference in $p(k)$ for a term t_i one step away from the term t is $\frac{5}{30}$ but for a t_j five steps away from term t, the weighting is $\frac{1}{30}$. If terms such as t_j are highly related to the term t then the strength of this relationship is severely degraded.

Alternatively, since we did not examine window sizes greater than five, it may be that a linear weighting scheme is much more important. This is probably due to the amount of risk involved with associating two terms k steps apart. The further the terms are apart the greater the risk in forming a legitimate association between them. Further analysis is required to determine whether this is the case and what the ideal window size is. Certainly, from these results the largest window size consistently outperformed the smaller window sizes, though this performance increase diminished as k increased.

We also examined two different methods to select quality properties (mean threshold as in the original version and the use of the Log Likelihood Ratio). The application of selection after the concept combination increased performance over selection of quality properties beforehand. When we compared the difference in performance between the mean threshold and the LLR, improved IR performance was achieved with the LLR method. However the result was marginal, approximately half a percent increase in mean Average Precision. Whilst this gain was not statistically significant, it may lead to further improvement when an information inference component is used to expand the query instead of the concept combination. An explanation as to why the performance increase was not as dramatic as we expected is that the terms that were removed using the LLR tended to be rather common terms used in that context. i.e terms that are relatively non-informative, so while removing such terms creates a cleaner representation of the context given the query the effect to performance is marginal because keeping the terms will not influence ranking dramatically. However, this may be problematic when performing the information inference component which relies on matching the contexts of terms which appear in a similar context as the query.

When adding terms to expand the query improved performance was achieved around 85 to 100 extra terms. When adding extra terms to the original query it appeared that once about a hundred terms were added to the query the IR performance began to plateau. Presumably, this is because the weighting assigned by adding another term was less than previously (since we added the top n terms to the query), thus further additional terms made less contribution to the overall ranking of the documents.

In a comparison between the baseline method against the original HAL-O space, the best HAL, pHAL-1 and pHAL-2 methods, we found that an improved retrieval performance was achieved which was statistically significant. Whilst the probabilistic variants did not significantly outperform the original methods, we showed that a competitive alternative can be developed in a probabilistic framework.

There are several differences between our proposed combination methods and the original method. Firstly, we perform a simple weighted linear combination of the concepts, without any re-weighting of specific terms or re-normalization. By not doing so the effect in terms of IR performance is minimal. However, the idea of disambiguating the context terms is a rather sensible notion. We attempted to perform disambiguation by adopting a simple technique from Amati [1] where only common terms in two or more contexts are kept. As a result the Method 2 combination technique obtained the best performance of all the combination methods. Potentially, there are other ways to determine the correct context given the query terms as previously mentioned in earlier sections.

There appears to be some problems with each of the methods with respect to this issue as they all depends on the usage of terms within the corpus. Firstly, if the majority of the query terms occur together in multiple contexts, then we may not have enough information to discriminate which is the correct context the query terms are used in. Secondly, if one particular context is stronger than another given the query terms then again the incorrect context is chosen.

An example: *Black Monday* ('black mondai') which refers to a stock market crash on the 17th of October, 1987. Given the process of concept combination the term 'black' is the dominant term as it has a higher Inverse Document Frequency score. This is the first problem incurred in this example with respect to disambiguating the context of the query terms. There are many contexts in which the term 'black' occurs, for instance, black voters, black Americans, black community, black and decker, black and white, etc, which are not in the same context as the stock market crash (see Table 7). After combination of the concepts for each method a mixture of terms exists within the list of the top twenty terms. After the LLR is employed very common terms in the collection are removed; they just so happen to be related to query (i.e. 'stock', 'market', 'share', etc.) however these terms are not very informative terms as many of the documents with in the collection contain these terms and the Inverse Document Frequency scores are very low. Nonetheless, the incorrect context is chosen regardless of method.

The reversal of terms for concept combination 'mondai' $\oplus$ 'black' results in a different set of top terms where the context is stock market oriented generally but not specifically about the topic 'Black Monday' (see Table 7).

An additional term however provides more information for the process of disambiguation and a more coherent set of expansion terms are generated (see Table 7). The difference in performance is significant. The query Black Monday ('black mondai') alone, fetches less than one percent mean Average Precision, and so do all the query expansion methods. However, if we add the term Brady ('bradi' - a key term in the report about the stock market crash) then the expansion is fruitful and a mean average precision of 30.1 and 27.2 for the HAL and HAL-LLR methods is obtained, respectively. In contrast issuing the three terms as a short query results in 22.2, while the human expanded query obtains 26.9 mean Average Precision.

This example is quite indicative of the difficulties involved in expanding a query. Again, if there are not enough query terms to sufficiently disambiguate the context of the query then the most dominant context will prevail. The ordering of the combination process determines the dominant context - changing the order affects the representation. Thus it is imperative to elicit the correct ordering for combination purposes or even change the representation (perhaps by adopting a tree

black	mondai	REL	HQE
decker	said	market	crash
white	1988	stock	market
black	short	trade	document
mr	new	futur	mondai
mondai	auction	report	black
sai	close	crash	stock
market	market	bradi	octob
said	1	exchang	19
american	trade	commiss	1987
vote	share	secur	reason
year	night	system	commiss
voter	8	specialist	futur
south	stock	recommend	exchang
compani	mr	sec	trade
new	price	limit	guard
hispan	tokyo	regul	analysi
commun	9	margin	contain
peopl	compani	oct	attempt
polit	term	octob	chicago
busi	rate	board	secur

Table 9. The top terms given the terms 'black' and 'mondai', the top terms given all the relevant documents for the topic (REL) and top terms given human query expansion (HQE).

structure). However, this may not actually benefit performance, as in the example above. Given sufficient context, the (p)HAL space is able to expand the representation of the query in a contextually coherent fashion which is comparable to and improves upon the human expanded query.

HAL	HAL-LLR	pHAL1	pHAL-LLR	pHAL2	pHAL2-LLR
black	black	black	black	black	black
mondai	mondai	mondai	mondai	mondai	mondai
decker	decker	decker	decker	decker	decker
white	white	white	white	white	white
mr	vote	mr	vote	mr	vote
sai	voter	said	short	said	short
market	south	market	auction	market	voter
said	hispan	sai	voter	sai	auction
american	commun	new	night	new	south
vote	polit	year	south	year	night
year	leader	compani	hispan	american	commun
voter	democrat	american	leader	compani	leader
south	famili	share	jackson	vote	jackson
compani	jackson	1	tokyo	share	republican
new	republican	close	republican	1	tokyo
hispan	standard	vote	student	close	student
commun	hole	1988	hole	stock	tuesdai
peopl	student	stock	popul	1988	popul
polit	school	short	township	short	women
busi	township	trade	colleg	voter	colleg

Table 10. List of the top twenty words given the concept combination of 'black' $\oplus$ 'mondai' and selection of quality properties.

8. Limitations

This initial study examines only the first three processes of the method introduced in [15, 3]. Substantially better results have been reported by making inferences after the concept combination. The information inferences are derived using Information Flow theory[2] where the query is expanded with semantically similar terms. Further work shall examine the extension to the framework.

Reversal HAL	Reversal HAL-LLR	Addition HAL	Addition HAL-LLR
mondai	mondai	bradi	bradi
black	black	mondai	mondai
said	1988	black	black
1988	short	said	task
short	auction	report	night
new	close	commiss	panel
auction	night	market	recommend
close	tokyo	new	read
market	term	mr	dillon
1	index	short	quot
trade	averag	close	circuit
share	yen	stock	baker
night	treasuri	trade	kei
8	tuesdai	treasuri	confirm
stock	trader	1	appoint
mr	morn	task	feb
price	yield	forc	amend
tokyo	nikkei	night	observ
9	wednesdai	8	weekend
compani	fridai	price	rumor
term	p	presid	conclud

Table 11. Reversal of concept combination Monday ('mondai') $\oplus$ Black ('black') and the combination of addition term Brady ('bradi') $\oplus$ Black ('black') $\oplus$ Monday ('mondai') .

Other limitations include the collection; for the purposes of analysis we have used only one collection the Wall Street Journal. The characteristics of this collection, i.e newspaper style articles on current affairs, limit the extent to which we can generalize about the aforementioned results. The size of vocabulary has been restricted by removing terms that only occur a few times, and built from only 40000 documents. Also, we have applied stemming (which has been noted in the literature to slightly degrade IR performance). The size of the window in our case maximum of 5 doubled to 10 because both directions are taken into account, whereas window sizes of up to 10(20) have been used. All these factors invariably will affect the quality of the (p)HAL spaces constructed as we have already witnessed in this small scale experiment.

9. Conclusions and Further Work

We have offered a probabilistic variant to that proposed in the original work. We have shown that the pHAL methods outperform the HAL methods given the task of query expansion using concept combination. However this work is preliminary and requires further work to be performed to confirm the improvement and to fully assess the utility of pHAL space. The following items will be considered for further work:

- Window Size and Weight - Perform analysis at higher window sizes and determine whether the weighting is influential on the quality of IR performance.
- Apply a more decision oriented approach to the combination of concepts in order to disambiguate the context of the query terms.
- Implement the inference component of the model.
- Whilst we have used this model for query expansion under a query likelihood approach, the expanded query representation could alternatively be used in a document likelihood approach (akin to relevance models).
- Application to a number of collections, using a larger term space and more queries.
- Examining a set of one term queries, so that we can remove the concept combination process and focus on the expansion to determine whether a better expansion is obtained with the different selection methods.

10. Acknowledgements

The author would like to thank Peter Bruza for his helpful comments through the course of this study and Malcolm Crowe for his comments and edits. This material is based on work supported in part by the University of Paisley and Memex Technology Ltd.(www.memex.com). Any opinions, findings and conclusions expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

References

1. G. Amati, *Divergence from randomness*, Ph.D. thesis, Department of Computer Science, University of Glasgow, 2003.
2. J. Barwise and J. Seligman, *Information flow: The logic of distributed systems*, Cambridge Tracts in Theoretical Computer Science, no. 44, 1997.
3. P. D. Bruza and D. Song, *Inferring query models by computing information flow*, Proceedings of the eleventh international conference on Information and knowledge management, ACM Press, 2002, pp. 260–269.
4. Burgess C., K. Livesay, and K. Lund, *Explorations in context space: Words, sentences, discourse*, Discourse Processes **25** (1998), no. 2-3, 211–257.
5. P. Gardenfors, *Conceptual spaces: The geometry of thought*, MIT Press, 2000.
6. D. Harman, *Relevance feedback revisited*, The 5th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Copenhagen, Denmark), 1992, pp. 1–10.
7. J. P. Levy and J. A. Bullinaria, *Learning lexical properties from word usage patterns: Which context words should be used*, Proceedings of thte Sixth Neural Computation and Psychology Workshop, Springer, 2001, pp. 273–282.
8. W. Lowe, *Towards a theory of semantic space*, Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society, 2001, pp. 576–581.
9. W. Lowe and S. McDonald, *The direct route: Priming in semantic space*, Proceedings of the Seventeen Annual Meeting of the Cognitive Science Society, 2000, pp. 600–665.
10. K. Lund and Burgess C., *Producing high-dimensional semantic spaces from lexical cooccurence*, Behavior Research Methods, Instruments and Computers **28** (1996), no. 2, 203–208.
11. C. D. Manning and H. Schutze, *Foundations of statistical language processing*, MIT Press, Cambirdge, Massachusetts, 2000.
12. M. Mitra, A. Singhal, and C. Buckley, *Improving automatic query expansion*, 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998, pp. 206–214.
13. G. B. Newby, *Congintive space and information space*, Journal of the American Society for Information Science **52** (2001), no. 12, 1026–1048.
14. M. F. Porter, *An algorithm for suffix stripping*, Program **14** (1980), 130–137.
15. D. Song and P. D. Bruza, *Discovering information flow using a high dimensional conceptual space*, The 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (New Orleans, LO), 2001, pp. 327–333.

16. C. Zhai and J. Lafferty, *A study of smoothing methods for language models applied to ad hoc information retrieval*, ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR) (Tampere, Finland), ACM Press, 2001, pp. 49–56.

c/o School of Computing, University of Paisley
E-mail address: leif.azzopardi@paisley.ac.uk