Australian Government
Department of Defence
Defence Science and
Technology Organisation

Comparison of Human and Latent Semantic Analysis (LSA) Judgements of Pairwise Document Similarities for a News Corpus

Brandon Pincombe

Intelligence, Surveillance and Reconnaissance Division
Information Sciences Laboratory

DSTO-RR-0278

ABSTRACT

Pairwise similarity judgement correlations between humans and Latent Semantic Analysis (LSA) were explored on a set of 50 news documents. LSA is a modern and commonly used technique for automatic determination of document similarity. LSA users must choose local and global weighting schemes, the number of factors to be retained, stop word lists and whether to background. Global weighting schemes had more effect than local weighting schemes. Use of a stop word list almost always improved performance. Introduction of a background set of similar documents increased larger correlations and reduced smaller ones. The correlations ranged between approximately 0 and 0.6 depending on the LSA settings indicating the importance of correct settings. The low maximum correlation indicates that information presentation schemes based on LSA may often be at variance with visualisations based on human decisions even using the best settings for a data set.

Published by

DSTO Information Sciences Laboratory
PO Box 1500
Edinburgh South Australia 5111 Australia

Telephone: (08) 8259 5555
Fax: (08) 8259 6567

AR-013-177
September 2004

APPROVED FOR PUBLIC RELEASE

Comparison of Human and Latent Semantic Analysis (LSA) Judgements of Pairwise Document Similarities for a News Corpus

Executive Summary

Numerous areas, such as document visualization tools, Internet news sites, search engines or library retrieval systems, require an accurate method of assessing document similarity in order to carry out clustering, classification or search tasks. This report compares the performance of a commonly used technique, Latent Semantic Analysis (LSA), with human similarity judgements and identifies the settings that lead to the highest correlations.

LSA is a modern and commonly used technique for automatic determination of document similarity. In LSA local and global weighting functions are applied to a term x document matrix. A "latent semantic space" is constructed through singular value decomposition followed by matrix reconstruction using only the largest of the singular values. LSA produces pairwise similarities between the documents. Previously, assessments of the similarity of the judgements of these techniques to those of people have been carried out on data sets where the documents have been assigned to topical groups by human assessors or have been assigned a quality score on a continuum by human judges. Here the intermediate stages required in previous studies are eliminated by using a corpus in which multiple human assessors rated the pairwise similarities of all non-self-similar pairs in a set of fifty documents. Through simply testing the correlation between human relevance judgements of pairwise similarity and machine judgements of pairwise similarity within the same document set the need for large numbers of documents marked-up with relevance judgements is removed. There is no need to train the system to achieve an appropriate clustering threshold as no clustering is performed.

Three global and three local weighting schemes are compared for correlation to the set of human pairwise judgements. Global weighting is shown to have more effect on the correlations than local weighting. Amongst the global weights entropy is marginally better than normalised and both are significantly better than IDF. The use of stop word lists and background documents in the construction of the latent semantic space are shown to produce significant increases in the correlation with human judgements for the better weighting schemes. The correlations between LSA and human judgements of pairwise document similarity varied between 0.5988 and –0.0144. This indicates that care must be taken in the implementation of LSA and that, even when optimal settings are used, the resulting similarities are far from a perfect reflection of human judgements.

Authors

Brandon Pincombe
Intelligence, Surveillance and Reconnaissance Division

Brandon Pincombe gained first-class honours in Applied Mathematics in 1992 at Wollongong University and graduated with a Ph.D. in Applied Mathematics from the University of Adelaide in 1999. In 1997 he moved from a position as a computer officer (HEO-5) at Adelaide University to commence employment as a Research Scientist with Secure Communications Branch at DSTO.

Contents

Introduction

Numerous areas, such as document visualization tools, Internet news sites, search engines or library retrieval systems, require an accurate method of assessing document similarity in order to carry out clustering, classification or search tasks. Various systems of measuring document similarity are available. This report compares the performance of a commonly used technique, Latent Semantic Analysis (LSA), with human similarity judgements and identifies the settings that lead to the highest correlations.

LSA is an automatic method that assesses pairwise similarities between documents. In LSA a term x document matrix is created and weighting functions are applied to this matrix followed by singular value decomposition. A "latent semantic space" is constructed through reconstituting the term x document matrix using only the largest of the singular values. LSA assumes there is a latent semantic structure in the documents it analyses and that the singular values are loosely associated with concepts. Various local and global weighting schemes are available for LSA. These alter the level of correlation with human judgements as does the choice of the number of singular values to use in the construction of the latent semantic space. The effects of the settings of these variables on the correlation of the pairwise document similarity judgements of LSA and human assessors are explored in this report.

Automated methods of measuring document similarity overwhelmingly measure the pairwise similarities between documents and LSA is no exception. However, assessments of the similarity of the judgements of these techniques to those of people have previously been carried out exclusively on data sets where the documents have been assigned to topical groups by human assessors or have been assigned a quality score along a continuum by human judges. The problem with such assessments is that an extra stage is required. These pairwise similarities are either used by a clustering algorithm to produce "topical" groups or the dimensionality of the pairwise judgements is reduced to a single dimension to give a continuum of judgements. This introduces problems in differentiating whether poor performance is due to the clustering or dimensionality reduction algorithms being used or the underlying technique being investigated. Further problems arise due to the inaccuracy and incompleteness of the topical groupings in the topically grouped corpora available[1]. These corpora are typically constructed to give very large numbers of documents with relevance judgements associated with them. Document classification is a highly subjective process but, in order to process the large numbers of documents required for accurate comparison of many methods, it is necessary to rely on a single assessor of topicality for the great majority of documents. The assessment process involves assigning a document to none, one or more of a large group of topics. Inevitably errors are made, for example a number of articles on Ivory Coast politics (not related to elephant ivory) appear under the elephant ivory trade topic in TREC. These problems

[1] Commonly used corpora include Reuters, Cranfield, Time, Medlars, adi, cacm, cisi, lisa, npl, TDT and TREC.

are avoided in this study by using a corpus in which multiple human assessors rated the pairwise similarities of all non-self-similar pairs in a set of fifty documents. This provides a basis of human judgements with which machine techniques can be compared. Through simply testing the correlation between human relevance judgements of pairwise similarity and machine judgements of pairwise similarity within the same document set the need for large numbers of documents marked-up with relevance judgements is removed. There is no need to train the system to achieve an appropriate clustering threshold as no clustering is performed.

Three global and three local weighting schemes are compared for correlation to the set of human pairwise judgements. Global weighting is shown to have more effect on the correlations than local weighting. Entropy global weighting is marginally better than normalised global weighting and both are significantly better than IDF weighting. The use of stop word lists and background documents in the construction of the latent semantic space are shown to produce significant increases in the correlation with human judgements for the better weighting schemes. The correlations between LSA and human judgements of pairwise document similarity varied between 0.5988 and –0.0144. This indicates that care must be taken in the implementation of LSA and that, even when optimal settings are used, the resulting similarities are far from a perfect reflection of human judgements.

Latent Semantic Analysis

Latent Semantic Analysis (LSA) (Deerwester, Dumais, Landauer, Furnas and Harshman, 1990) is a technique for inferring the contextual-usage meaning of words through a reduced dimensionality representation of documents in the corpus of interest. The aggregate of all the contexts in which a given index term, typically a word, appears or does not appear provides a set of mutual constraints that essentially determines the similarity of meaning of index terms to each other (Deerwester et al., 1990). LSA induces these semantic similarities in terms through dimensionality reduction to form a latent semantic space (Hoffman, 1999). This is achieved by singular value decomposition followed by reconstruction using a limited set of eigenvectors with the largest eigenvalues. The use of semantic correlations based on the evidence across the entire corpora reduces the problem with polysemy[2], synonymy[3] and

[2] A polysemous word has a diversity of meanings. For example right can mean the side turned east when facing north or the privilege of stockholders to subscribe to additional share issues at an advantageous price or in accordance with what is good, proper or just or politically conservative or having an axis perpendicular to the base or prompt and immediate or conformity with fact …

[3] Synonyms are two or more words sharing a meaning that is the same or nearly the same, for example solution and answer.

inflexion[4] inherent in purely term-based representations of documents (Landauer, 1997).

To reduce the effects of or eliminate polysemy, synonymy and inflexion on document similarity measures LSA relies on the use of an appropriate number of factors to reconstruct the term document matrix following singular value decomposition (SVD) (Berry, 1992). While there are theoretical guides to selecting the number of factors based on the rank-plus-shift method (Zha and Zhang, 1998) and probabilistic models relating frequency statistics to an underlying distribution (Papadimitriou, Raghavan, Tamaki, and Vempala, 1998; Hoffman, 1999) the empirical studies of retrieval performance (Deerwester et al., 1990; Dumais, 1990; Landauer, 1997; Landauer and Dumais, 1997) are probably a better guide in selecting the optimal number of factors. It is evident that low numbers such as seven (Koll, 1979) or 21 (Borko and Bernick, 1963) are too few. The SVD process does not allow for more dimensions to be used than there are documents in the corpus; so in the case investigated in this report there is either a restriction to 50 dimensions or a need for the use of backgrounding documents to raise the number of dimensions available for LSA. Empirical work has shown that the number of dimensions required depends on how topically focused the text under investigation is (van Rijsbergen, 2002). In work on a set of abstracts for medical papers (Deerwester et al., 1990; Dumais, 1990) in the order of 100 dimensions was found to be optimal. On more diverse document sets the use of 100 to 300 factors seems to give optimal performance (Landauer, 1997) under some circumstances and of about 300 to 325 (Landauer and Dumais, 1997; Landauer, Laham, and Foltz, 1998) in others. In practice terms and documents are typically represented by 200 to 300 of the largest singular vectors and then matched against user queries because of the computational requirements for managing LSI-encoded databases (Berry, Dumais and O'Brien, 1995).

During LSA the d text documents are placed in the columns of a matrix **N** with each row representing one of the t unique index terms. Therefore **N** is a txd matrix. Typically index terms are individual words, without any stemming applied. The next step in LSA is not SVD but the application of local weighting, expressing the importance of a word within a document, followed by global weighting, scaling by the degree to which the word carries information in the domain of discourse, to the matrix **N** (Deerwester et al., 1990). The details of the weighting schemes used in this report are discussed below in section 2.1. If factor analysis were undertaken at this stage instead of SVD, a dxd cosine similarity matrix **S** would be formed by taking the dot product of **N** and its transpose, i.e. $\mathbf{S} = \mathbf{N} \cdot \mathbf{N}^T$. The eigenvalues[5] $[\lambda_1, \ldots, \lambda_d]$ and corresponding

[4] Inflexion is the process or device of adding affixes to or changing the base form of a word to express syntactic function without changing its form class. Some examples are undone, done, doing, did, doer, do or follow, follower, following, follows.

[5] The eigenvalues are, by convention, ordered from largest to smallest. Eigenvalues are sometimes called characteristic roots or latent roots. The eigenvalue with the largest absolute magnitude is the spectral radius.

eigenvectors[6] [$\mathbf{x_1}$, …, $\mathbf{x_d}$] of the matrix **S** would then be found. SVD is more complicated than factor analysis but the concept of eigen-analysis must be kept in mind to understand the composition of the matrices formed by SVD. Subsequent to weighting SVD is performed on the matrix **N** to determine the matrices **U**, **L** and **V** such that $\mathbf{N} = \mathbf{ULV}^T$. The details of these three matrices are:

- The txm[7] matrix **U** describes the original row (or index term) entities as vectors unit length composed of derived orthogonal factor values[8]. Within this paper m=d, as the number of terms far exceeds the number of documents, so **U** is a txd matrix. The matrix **U** is related to factor analysis by being the matrix of eigenvectors of the square symmetric matrix $\mathbf{S} = \mathbf{N} \cdot \mathbf{N}^T$ (Deerwester et al., 1990).The mxd matrix **V** describes the original column (or document) entities in the same way as **V** describes the row (or index term) entries. In practice **V** is dxd. The matrix **V** is related to factor analysis by being the matrix of eigenvectors of the square symmetric matrix $\mathbf{S} = \mathbf{N}^T \cdot \mathbf{N}$. (Deerwester et al., 1990).
- The mxm diagonal matrix **L** contains the scaling values or singular values such that when the three components are matrix-multiplied, the original matrix is reconstructed. As it is unlikely the number of documents exceeds the number of terms in practice **L** is a dxd matrix. The values contained along the diagonal of **L** are the square roots of the magnitudes of the eigenvalues found in factor analysis, i.e. the diagonal elements of **L** are $\{\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d}\}$ (Deerwester et al., 1990).

When fewer than the m (=d in this report) factors found in SVD are used in reconstruction the resulting matrix is a least-squares best fit. One can reduce the dimensionality of the solution simply by zeroing all but the largest k (k≤d) coefficients in the diagonal matrix. This process produces a latent semantic space $\mathbf{N_k} = \mathbf{U_k L_k V_k}^T$ where only the first k eigenvalues of **U** and **V,** and the first k singular values of **L** are used in the reconstruction.

Weighting functions

The matrix $\mathbf{N} = [n_{ij}]$ is transformed by local and global weighting functions to make the cell contents better approximations of the interrelations between terms and documents (Nakov, Popova and Mateev, 2001; Witter, 1997; Dumais, 1991; Deerwester et al., 1990). Thus, the transformed matrix $\mathbf{N^*} = [n_{ij}^*]$ is formed on a term-by-term basis such that $n_{ij}^* = n_{ij}\, w_{ij}$ where the total weight w_{ij} of term i in document j is the product of the local and global weights given by:

$$w_{ij} = L(i,j) \times G(i). \qquad \ldots(1)$$

[6] The eigenvectors or characteristic vectors are linearly independent components or factors of the original dxd matrix S. The set of all eigenvectors is called the spectrum of S.
[7] The value of m is the rank of N and thus is less than or equal to the smaller of t and d.
[8] As **U** and **V** are composed of orthogonal vectors of unit length they are orthonormal matrices.

Here, L(i,j) is the local weight of term i within document j, and G(i) is the global weight of term i across all documents in the corpus (Nakov, Popova and Mateev, 2001; Witter, 1997; Dumais, 1991; Deerwester et al., 1990).

Local Weighting

Three local weighting functions (Dumais, 1991; Sparck-Jones, 1972) are investigated by varying the weight they place on common terms from 'term-frequency' (highest) through the base-2 logarithm of term frequency to binary weighting. The term-frequency is determined by the formula

$$\mathrm{tf}(i,j) = \frac{c(i,j)}{\sum_{k=1}^{I} c(k,j)} \quad , \qquad \ldots(2)$$

where tf(i,j) is the term frequency of the i-th term in the j-th document, c(i,j) and c(k,j) are the counts of the number of appearances of the i-th and k-th terms in the j-th document and I is the total number of terms. Given this definition, the term-frequency weight function is given by

$$L(i,j) = \mathrm{tf}(i,j)\,, \qquad \ldots(3)$$

the logarithmic weight function is given by

$$L(i,j) = \log_2\left(\mathrm{tf}(i,j) + 1\right), \qquad \ldots(4)$$

and the binary weight function by

$$L(i,j) = \begin{cases} 1 & \text{when} \quad \mathrm{tf}(i,j) > 0 \\ 0 & \text{when} \quad \mathrm{tf}(i,j) = 0 \end{cases} \quad . \qquad \ldots(5)$$

Global Weighting

Numerous global weighting functions have been suggested for LSA of which the most common are entropy, inverse-document-frequency, normal, none, real-entropy and global-frequency-inverse-document-frequency. The first three of these are used in this paper. To carry out no weighting at all a global weighting function of $G(i) = 1$ is used. Defining df(i) as the fraction of documents containing the i-th term (i.e. the document frequency of the i-th term); gf(i) as the corpus-wide global frequency for the i-th term and the conditional probability p(i,j)=tf(i,j)/gf(i,j) where there are i=1…I terms and j=1…J documents, it is possible to write the normal global weight as

$$G(i) = \frac{1}{\sqrt{\sum_{j=1}^{J} L(i,j)^2}}; \qquad \ldots(6)$$

the inverse document frequency (idf) global weight as

$$\mathrm{G(i)} = 1 + \log_2\left(\frac{\mathrm{J}}{\mathrm{df(i)}}\right); \qquad \ldots(7)$$

the entropy global weight function as

$$\mathrm{G(i)} = 1 + \frac{\sum_{\mathrm{k=1}}^{\mathrm{J}} (\mathrm{p(i,k)} \cdot \log_2 \mathrm{p(i,k)})}{\log_2 \mathrm{J}}, \qquad \ldots(8)$$

or

$$G(i) = 1 - \frac{H(d|i)}{H(d)} \qquad \ldots(9)$$

where $H(d|i) = -\sum_{\mathrm{k=1}}^{\mathrm{J}} (\mathrm{p(i,k)} \cdot \log_2 \mathrm{p(i,k)})$ is the entropy of the conditional distribution given i and $H(d) = \log_2 J$ is the entropy of the document distribution. Global-frequency-inverse-document-frequency (GfIdf) and real-entropy global weights have been used in the literature but are not considered in this work. The GfIdf global weight is defined by

$$\mathrm{G(i)} = \frac{\mathrm{gf(i)}}{\mathrm{df(i)}}, \qquad \ldots(10)$$

and is the only global weighting function to consistently perform more poorly than no global weighting at all in analysis of English literature texts (Nakov et al., 2001). The real entropy of the conditional distribution $G(i) = H(d|i)$ is less commonly used and is similar to the entropy global weight function. For a given local weighting function the use of real-entropy, entropy and inverse-document-frequency global weights seem superior for the analysis of English literature texts to normal and global-frequency-inverse-document-frequency weights or no weight at all in all cases although the exact order within these two groupings are variable (Nakov et al., 2001).

Document Similarity

To determine the similarity of a query to the documents in the (k x d) reduced rank latent semantic space, **N**, the query needs to be represented in that space, i.e. as a k length vector, and then compared to the d documents using a similarity measure. While a number of similarity measures, such as the correlation, Jaccard and Overlap

coefficients, are available it is standard practice within the LSA literature to use the cosine coefficient:

$$\mathrm{sim}(N_j, q) = \frac{N_j^T q}{\|N_j\|_2 \|q\|_2} = \frac{\sum_{i=1}^{k} (N_{ij} q_i)}{\sqrt{\sum_{i=1}^{k} N_{ij}^2} \cdot \sqrt{\sum_{i=1}^{k} q_i^2}} \quad \ldots(11)$$

where N_j is the kx1 reduced rank representation of the jth document and q is the kx1 query vector.

Uses of LSA

LSA assumes there is a latent semantic structure[9] in the documents it analyses. LSA uses a reduced SVD form to model the latent semantic structure in the standard term-document matrix that is normally obscured by noise or variability in word usage. Its initial use (Deerwester et al., 1990), and the use to which it is put here, was in determining document similarity. Educational applications of LSA (Laham and Landauer, 1998) include categorising text passages by textual complexity, automatically grading essays and as a text summarization tool (Laham and Landauer, 1998; Steinhart, 2001). LSA has also been used as a theoretical model of the acquisition and representation of knowledge in human memory (Landauer and Dumais, 1997)[10].

When appropriately trained LSA provides identical retrieval rates in a bilingual corpus for queries in either language on documents in either language (Landauer and Littman, 1990; Dumais, Landauer and Littman, 1996; Littman, Dumais and Landauer, 1998; Rehder, Littman and Dumais, 1998). Documents in one language are paired with their translations and this is treated as a single document for training purposes. The TREC-6 standardized cross language information retrieval competition (Schaüble and Sheridan, 1998) showed basic LSA (Rehder, Littman, Dumais and Landauer, 1998) produces inferior precision in an information retrieval task than similarity thesauri (Mateev, Munteanu, Sheridan, Wechsler, and Schäuble 1998; Sheridan and Ballerini, 1996), dictionary look-up (Hull and Grefenstette, 1996; Gaussier, Grefenstette, Hull, and Schulze, 1998; Allan, Callan, Croft, Ballesteros, Byrd, Swan and Xu, 1998; Davis and Ogden, 1998) and a generalized vector space model (Carbonell, et al. 1997). In mitigation, the LSA entry simply concatenated the documents and automatically produced a cross-lingual retrieval system. Cross language homonyms, particularly names and numbers, were not used to improve performance as they could have been and were in other systems. Interestingly, LSA performed better on non-English queries of English documents than on English queries of the same documents; better on French queries of French documents than on non-French queries of the same documents, and significantly better on any sort of querying of French documents than on English or German. Other methods always performed better when querying using the same

[9] Latent semantic structure can be thought of as "meaning".
[10] The rate of growth in picking word synonyms using LSA with 300 dimensions trained of a corpus of 30000 encyclopaedia entries matched that of children.

language as the document and always performed better on English than French. LSA provides an easily implemented solution to cross-language information retrieval with a lower precision than is obtainable using specially tailored schemes.

Related Work

In this report the pairwise similarity judgments rendered by LSA on a set of 50 documents are compared with those of human assessors. The effects of various alterable LSA parameters, such as local and global weighting schemes and number of factors, on the correlation with the human judgments are explored. Therefore, literature in which the performance of LSA is compared to that of human subjects or in which the effect on altering the weighting schemes or number of factors is of particular relevance.

Meaning Divorced from Word Order?

The level of meaning that can be extracted from English texts using LSA, i.e. essentially ignoring word order, has been explored to some extent (Landauer et al., 1997; Rehder et al., 1998; Wolfe et al., 1998). This was done by getting two humans to rate the knowledge in these texts and then comparing the inter-rater rater reliability and predictive accuracy of their estimates with those obtained from two LSA based methods of assessing the knowledge content of the text (Landauer et al., 1997). Little difference was found between the human and machine assessments.

Essays of 250 words on heart function and anatomy were written by 94 undergraduates and presented to two professional readers who agreed on a rating scheme reflecting quality and then independently rated each essay on a scale of 1 to 5. The students were also given a short answer test to rate their topical knowledge (Wolfe et al., 1998). LSA was trained on 803 sentence length passages[11] from 27 encyclopedia articles on heart function and anatomy producing a 94-dimensional latent semantic space. The pairwise cosine distances between all 94 documents were calculated in this space. Each document was assigned a score equal to the average of the ten most similar documents by the cosine measure weighted by their degree of similarity. The correlation of these scores with those of the human assessors was 0.77; the correlation of the LSA assigned scores with the short-answer test scores was 0.81 and the correlation of the human assessors with the short answer test scores was 0.70. The 94 essays were also scored by LSA by assessing their cosine distance from a textbook passage on heart function and anatomy. Using this method the correlation with human assessors was 0.72 and with the short answer test scores was 0.77 (Landauer et al., 1997).

A total of 273 undergraduate students were given 10 minutes to write an essay on one of three subjects[12] and essays were assessed by two assessors, one more experienced in

[11] The documents contained 3034 unique words. A stop list of 439 common words was used and the remaining 2595 words were used as the terms in the 2595x803 term-document matrix.

[12] The three subjects were attachment in children (55), aphasias (109) and operant conditioning (109). The numbers in brackets represent the number of students who wrote on each topic.

the area than the other but both more experienced than the students. LSA was trained on the 4904 paragraphs and 19153 unique words in the textbook used in the course the students were doing. No stopping was performed and the number of factors was varied between 2 and 2100 to maximize LSA performance. The growth in performance leveled off between 400 and 500 dimensions and became negative after 1500 dimensions. When the human evaluated quality scores from the ten nearest documents as assessed by the cosine measure were weighted-averaged according to their level of similarity to produce a score for each document the correlation with human assessors was 0.65 whereas the correlation between human assessors was 0.64 (Landauer et al., 1997).

This study used two LSA based methods to rate the knowledge contained in documents. The first relied on the accuracy of human raters as it determined the knowledge content through a weighted average of human assigned scores for documents assessed as similar using LSA and a cosine measure of similarity. This is rather self-referential and the good correlations with human judgments are not surprising. The second method assessed the knowledge in the texts using LSA by measuring their cosine similarity to a well-regarded reference text on the precise topic the texts. This second method of document rating was only used on one of the two document sets raising questions about whether it works as well in general or is reliant on particular features of the document set it was applied to such as a tight topical focus.

The most significant difference between this study and the present report is that the human raters appraised the perceived information quality of the documents rather than the pairwise inter-document similarities.

LSA prediction of learning rates

To the 94 psychology students, writing the 250 word essays in the study mentioned above, were added 12 medical students. LSA was trained on 36 articles containing 17880 words and 3034 unique words and a 100 dimensional space was constructed.

Effects of weighting on accuracy of a classification task

The effects of the different methods and combinations of weighting available in LSA on the accuracy of a classification task have been explored (Nakov et al., 2001). *The Adventures of Sherlock Holmes* and *Huckleberry Finn* were, respectively, broken into 272 and 269 two-kilobyte texts and the similarity of 245 and 242 of these to 54 withheld reference texts, 27 from each book, was measured using LSA. These text chunks were then assigned as part of one book or the other based on their similarity to the withheld text chunks of known source. The actual origin of the text chunks was compared with the predicted origin to come up with a measure of the accuracy of this technique for a number of combinations of local and global weighting.

The local weighting functions investigated were term-frequency and logarithmic (see equations (3) and (4) respectively). The global weighting functions considered were

normal, inverse-document-frequency, entropy (see equations (6), (7) and (8) respectively) as well as no global weighting, global-frequency- inverse-document-frequency and real entropy.

A stop list was used but no stemming or phrase identification was performed. LSA retained 15 dimensions based on the average point at which the singular values began to asymptote for the twelve local and global weighting combinations. This may have introduced problems as the optimal number of factors needed to balance loss of information (too few) and introduction of noise (too many) may not have been chosen for all combinations.

The precision of the results for the combinations of local and global weighting functions differed between the two texts. The general conclusions on the quality of the global weighting functions were that no weighting and global-frequency- inverse-document-frequency weighting consistently produced the least precise results; normal weighting produced good precision results with term-frequency local weighting on *Huckleberry Finn* but not on *The Adventures of Sherlock Holmes* or with logarithmic local weighting; inverse-document-frequency produced good precision results under all four conditions and one of entropy and real entropy led to the best results for both local weights and both texts.

This study (Nakov et al., 2001) considered the problem of text categorisation, which is different to the problem considered in this report of correlating human and machine similarity judgements. The present report noted the poor performance achieved by using no global weighting or global-frequency- inverse-document-frequency global weighting and did not implement them. Real entropy global weighting is reasonably similar to entropy global weighting in both formulation and performance and is not included in many commercially available LSA packages so, even though it slightly outperformed entropy weighting under the conditions reported in the literature, it has not been considered in this report.

Further text classification work

Nakov and associates have performed a number of studies classifying texts by genre, author and work (Nakov 2000a; Nakov 2001a; Nakov 2001b; Nakov et al., 2001). In separate studies works in Russian, Bulgarian, German and English have been broken down into chunks and the distances between these chunks have been measured using LSA. The interesting feature of these studies for the work in hand is that the optimal choice of the number of factors retained after SVD depends on the level of classification desired. Small numbers of factors classify by genre (poetry or prose), larger numbers by author and larger numbers by work. This challenges the concept that there is a generally optimal number of factors that must be retained for all classification tasks.

Documents

Human assessed pairwise similarities between the fifty documents contained in Appendix A1 are used in this report as the ground-truth to which LSA pairwise similarity assessments are compared. The experiment giving these similarities was carried out by a team led by Dr Michael Lee at the Psychology Department at Adelaide University as part of a DSTO contract. The documents were selected from a group of articles taken from the Australian Broadcasting Corporation's news mail service, providing text e-mails of headline stories, from August 2002. They were chosen so that each "topic" had an identifiable "sub-topic" within it, for example the articles on Iraq contained a grouping on the suicide of Abu Nidal and those on Australian politics contained a grouping on leadership ructions in the Democrats. The selection of sub-topics within topics was specified in the original contract in an attempt to ensure a broader spread of human judgements of document similarity. Assessors rated the similarity of document pairs on a scale between 1 [least similar] and 5 [most similar]. These similarity judgements were subsequently normalised onto a [0,1] scale. The similarity of each document pair was judged by an average of ten[13] different assessors. The use of multiple assessors of similarity for each document pair not only improved confidence in the mean similarity measure but also allowed an assessment of the level of confidence through examination of the variance and spread of scores.

Given that the fifty documents in Appendix A1 are the ground-truth that LSA pairwise similarities are compared to it is important to demonstrate that they lay within the bounds of typical English language documents. Similarly the 364 and 4172 ABC newsmail documents used to produce a background in LSA should be typical of English text. A cursory inspection shows that they are not excessively aberrant and this is confirmed by the word frequency spectrum in Figure 1 and the vocabulary growth rate in Figure 2. The word frequency spectrum in Figure 1 shows a plot of the actual frequencies of words appearing between one and 15 times in the document corpus (drawn as circles) and the expected values according to Sichel's generalised inverse Gauss-Poisson model (Sichel, 1975; Sichel, 1986; Sichel, 1997; Baayen, 2001, pp 89-93). Similarly, in Figure 2, the expected vocabulary growth, i.e. the number of unique words expected within a set of words, is shown as a line and the actual vocabulary growth is shown as circles.

[13] A small number of pairwise similarities were judged nine or 11 times but the average was precisely ten.

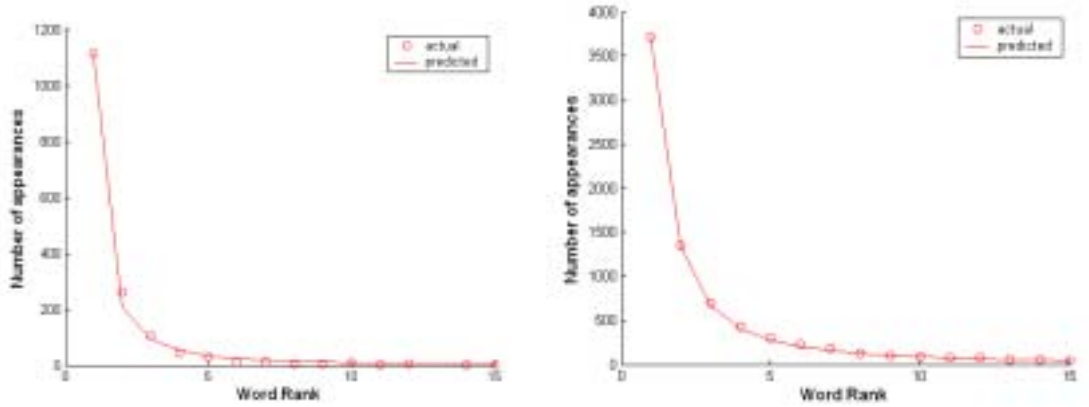

Figure 1: Expected and actual frequency spectrum for 50 and 364 ABC newsmail documents.

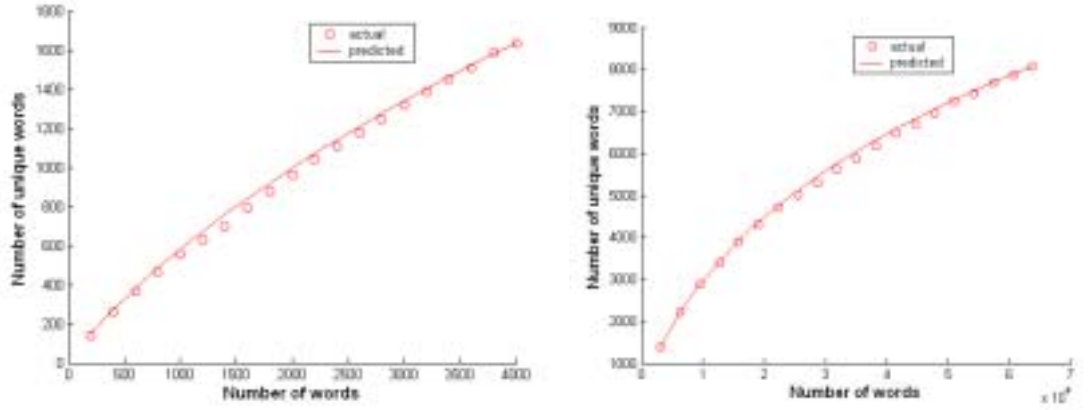

Figure 2: Expected and actual vocabulary growth in the 50 and 364 ABC newsmail documents.

The null hypothesis that there is no difference between the actual frequency spectrum and the expected frequency spectrum from five language models is tested using χ^2. In Table 1 the degrees of freedom, calculated X^2 statistic and p value from the χ^2 test are given for each of these five models over the set of 50, 364 and 4172 ABC newsmail documents as well as reference figures for *Alice in Wonderland*, *Through the Looking-glass*, *The War of the Worlds*, *Hound of the Baskervilles* and the context governed sub-corpus of the British National Corpus (BNC) as calculated by Baayen (Baayen, 2001, pp 126). The total number of words, N, and total number of unique words, V(N), are listed at the top of Table 1 for the three document sets used in this report and five comparison documents. Differing models of language are closest to different texts, for example the Yule-Simon model (Simon, 1961; Chen and Leimkuhler, 1989) is closest to Alice's Adventures in Wonderland and War of the Worlds, the extended Zipf's law (Khmaladze and Chitashvili, 1989) to Through the Looking-glass, the lognormal model (Carroll, 1969; Herdan, 1964) for Hound of the Baskervilles and the generalised inverse Gauss-Poisson model (Sichel, 1975; Sichel, 1986; Sichel, 1997; Baayen, 2001, pp 89-93) for the context-governed sub-corpus of the British National Corpus (Baayen, 2001, pp

124-131). As the p-statistics are uniformly low the null hypothesis can be rejected in all cases. Therefore all five of these commonly used language models fit the text of the 50 document and 364 document sets. The document sets used in this report can be regarded as well within the normal range of English language texts.

	50 Docs	364 Docs	4172 Docs	Alice…	Through…	The War…	Hound…	BNC
N	4000	63847	616675	26505	28767	59938	59241	6154206
V(N)	1636	8080	26376	2651	3085	7112	5741	79833
Lognormal model								
$X^2(14)$	45.32	26.17	2398.43	34.89	21.58	50.39	33.84	4424.22
p	0.0000249	0.02464	0	0.0015	0.0877	5.3E-6	0.0022	0
MSE	43.58	105.42	20866.94	100.43	53.62	148.94	106.05	52023.2
Generalised inverse Gauss-Poisson, γ=-0.5								
$X^2(14)$	-421.16	1233.48	136.86	262.06	302.65	1763.65	1322.82	567.96
p	0	0	0	0	0	0	0	0
MSE	1399.50	4778.13	3514.50	421.42	447.87	5672.58	3341.28	35770.77
Generalised inverse Gauss-Poisson, γ free								
$X^2(13)$	184.07	349.71	469.90	262.06	155.96	364.38	1317.06	1407.62
p	0	0	0	0	0	0	0	0
MSE	222.359	1337.76	8165.19	416.42	243.93	1108.5	3366.21	73032.73
Extended Zipf								
$X^2(15)$	127.75	989.60	794.12	29.05	22.22	19.74	227.84	121917.9
p	0	0	0	0.0158	0.1021	0.1822	0	0
MSE	1300.10	12052.51	82498.87	126.22	75.27	363.47	3925.04	11059532

Table 1: Statistics on ABC Newsmail documents and comparison texts.

Method

Infoscale Tools[14] subroutines were used for parsing text documents, forming and weighting term by document matrices, performing singular value decompositions on these matrices, reconstructing a term by document matrix of lower dimensionality and querying using this matrix. Four cases were considered. In each case the correlation between human judgements of pairwise similarities between the 50 documents in Appendix A1 and the similarities between these documents found using LSA was found. Self-similar documents were excluded from this process. In the first two cases just the 50 documents in Appendix A1 were used to construct the term-document matrix that then underwent singular value decomposition and was reconstructed using the top 10 to 300 factors. Initially this was done without eliminating any stop words from the documents and was then repeated using the stop word list in Appendix A2.

[14] Infoscale tools are a set of scripts written by Scott Deerwester, Pat Hawley, Sue Dumais and Steve Uhler to perform the constituent parts of LSA.

In the latter two cases this process was repeated using both the original 50 documents and another 314 documents to form the term-document matrix. These 314 documents were used as a background and were from a similar source being all ABC newsmail stories from December 2001. Again both lemmatised and un-stopped texts were processed and the results recorded. The details of the use of the Infoscale Tools scripts are described in the remainder of this section.

A file named "docs" containing the 50 documents (see Appendix A1) for which pairwise similarity judgements were wanted was established with each story separated by a single blank line. The query file "query.text" was constructed from the document file "docs" using the "mkey" command which parses text into a form dictated by its parameters. The form of the "mkey" command depended on whether stopping was required or not. The stop-words given in Appendix A2 were contained, one word per line, in the file "ccw" (short for closed-class-words). To construct a non-stopped query file the command "mkey –k10000 –l2 –m20 –M100000 docs > query.text" was used and for stopped queries the command was "mkey –k10000 –c ccw –l 2 –m 20 –M100000 docs > query.text". In both cases the "-k10000" allowed 10000 unique words to be used; "-l 2" eliminated words of length less than 2 ; the "-m 20" truncated words of length greater than 20 and the "-M100000" allowed a maximum line length of 100000 characters. The "-c ccw" command indicated that stop-words should be taken from the file "ccw". In this report the settings were such that all non-alphabetic characters were transformed into a single space, all alphabetic characters were transformed to lower case, all single letter words were removed and all words longer than 20 characters were truncated to 20 characters. Each of the resulting 50 queries in "query.text" was a parsed document from the document set and the similarity between these 50 documents of interest and all other documents was found later, although only the similarities with the 50 documents of interest were recorded. The file containing the relevance judgements necessary to generate the receiver operating characteristic (ROC) curves was then constructed in accordance with the style used in SMART. As the ROC curves were not used in this study the relevance judgements entered were pure place fillers simply indicating that each query in "query.text" had perfect relevance to the corresponding document in "docs". Future work may include entry of the human judgements of pairwise similarity into the relevance judgement file to produce ROC curves. If backgrounding was used the document files "docs" then had the additional 314 ABC news stories appended to the end of it.

In all cases the "pindex" command was used to construct the reduced dimensionality term-document matrix for the documents. As the "pindex" command calls "mkey" to parse the text, "weight" to apply the local and global weights and "las2" to perform singular value decomposition and then reconstruct the matrix with the principal factors it is necessary to pass it a considerable amount of information. When stopping was in use the form of the command was "pindex –c ../ccw –l 2 –m 20 –p /usr/local/tas/bin/las2 –n 100 –M 10000 –w tf idf docs" to use the stop list called "ccw" contained in the directory above, ignore words of length less than two, truncate words more than twenty characters long, pass across the full path to "las2" for singular

value decomposition, use 100 factors for reconstruction of the term-document matrix and use term frequency local weighting and inverse document frequency global weighting on the file docs.

The "Q" command was used to obtain a similarity measure between an individual query contained within "query.text" and each of the documents represented in the latent semantic space. This command was of the form "Q 1 –n 100 –s cos" where this exact command would have queried all the documents in "query.text" against the first document using 100 factors to do so and using the cosine similarity measure to assess the similarities between the queries and the document. In the two cases where backgrounding was used only the first 50 of these documents, corresponding to the 50 queries, were of interest as only they had experimental data for the human judgements of pairwise similarities.

Results and Discussion

The results given in Figure 3 through Figure 6, and in tabular form in Appendix B, show the correlation coefficient between various implementations of LSA and the human judgements mentioned earlier.

In Figure 3 the correlations between human pairwise similarities and those calculated using LSA on the 50 documents without any backgrounding or stopping are shown. The outstanding feature of Figure 3 is the much lower correlations between human judgements and LSA with IDF global weighting than with either entropy or normal global weights. There is a substantial increase in agreement between LSA and human judgments when the number of factors is increased from 10 to 20 for both entropy and normal global-weights. The use of logarithmic and term-frequency local weights combined with entropy global weighting gives the best results in Figure 3. Conversely, the use of logarithmic and term-frequency local weights with IDF global weighting gives the worst results. A maximum of 50 factors is displayed in Figure 3 because with only 50 documents used to generate the term-document matrix this number of factors represents no dimensionality reduction at all. It is interesting to note that the optimal performance of LSA is achieved at the point where there has been no dimensionality reduction.

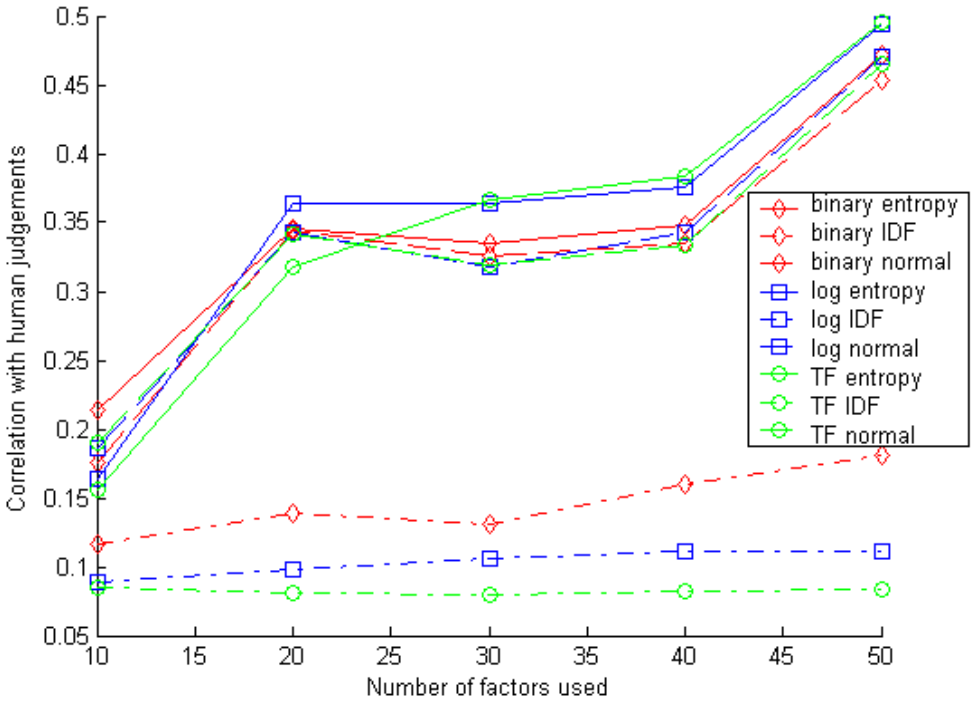

Figure 3: Correlation between unstopped LSA and human judgements.

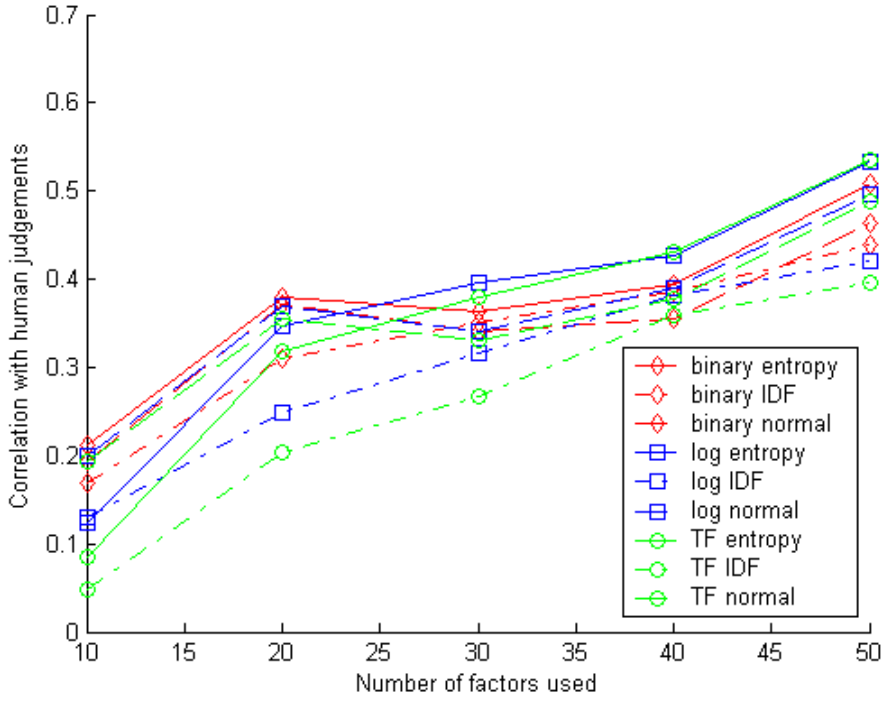

Figure 4: Correlation between stopped LSA and human judgements.

In Figure 4 the correlations between human and LSA pairwise similarity decisions are shown when LSA is performed on a document set with the stop words contained in Appendix A2 removed from it but without the use of backgrounding documents. The outstanding difference with Figure 3 is in the relative performance of LSA using IDF global weighting. When all 50 factors are used the correlations with human judgements for LSA using IDF global weighting are only slightly lower than those for the other two global weighting schemes. When ten factors are used the binary-IDF weighting combination shows a higher correlation with human judgements than either of the log-entropy or TF-entropy weighting combinations and log-IDF is superior to the latter.

Again there is a substantial increase in agreement between LSA and human judgments when the number of factors used for querying is increased from 10 to 20. There are a number of odd points with a binary-normal weighting combination having a local peak in similarity at 20 factors that is surpassed by 50 factors and for tf-normal weighting the correlation in results is better when 20 factors are used than when 30 factors are used. Apart from this the situation is as expected with an improving performance as the number of factors used increases up to the point where the number of factors equals the number of documents from where there is no improvement. The peak results are uniformly better when stopping is used than without stopping. This is particularly evident when idf global weighting is used but it is still the worst global weighting function in both cases.

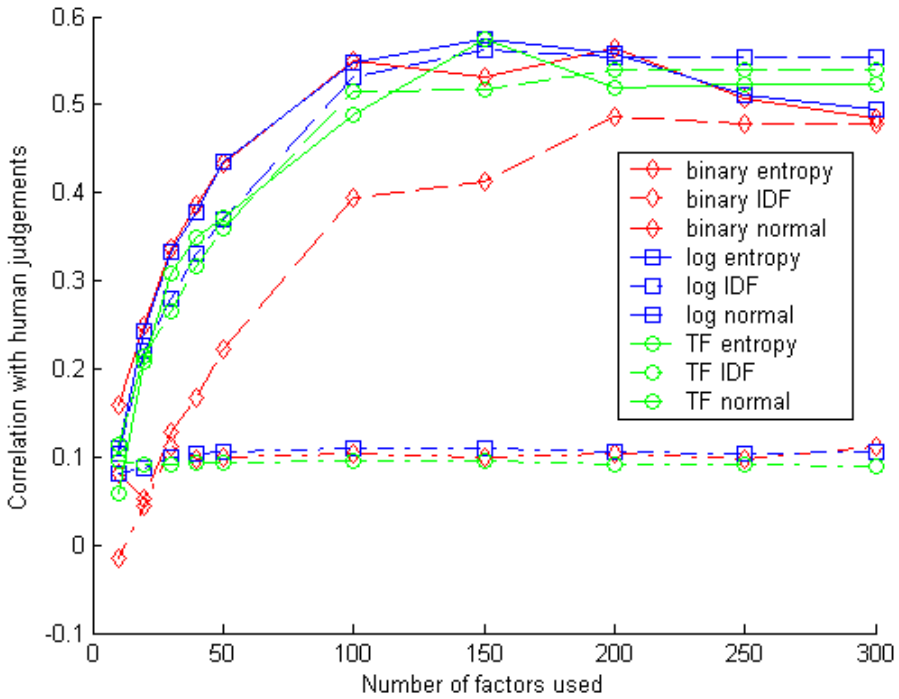

Figure 5: Correlation between unstopped, backgrounded LSA and human judgements.

The introduction of extra documents to provide a background against which LSA can work produced a reduction in performance for low numbers of factors but allowed performance to continue to improve as the number of factors grew towards the total number of documents. As seen in Figure 5 the level of correlation with human judgments at 300 factors was always greater than the non-stopped but un-backgrounded situation (Table 2). When compared with the results in Figure 4 it can be seen that the at 300 factors the unstopped but backgrounded LSA is superior to the stopped but un-backgrounded LSA for idf and normal global weights, inferior for entropy global weighting and has a greater maximum correlation with human judgments.

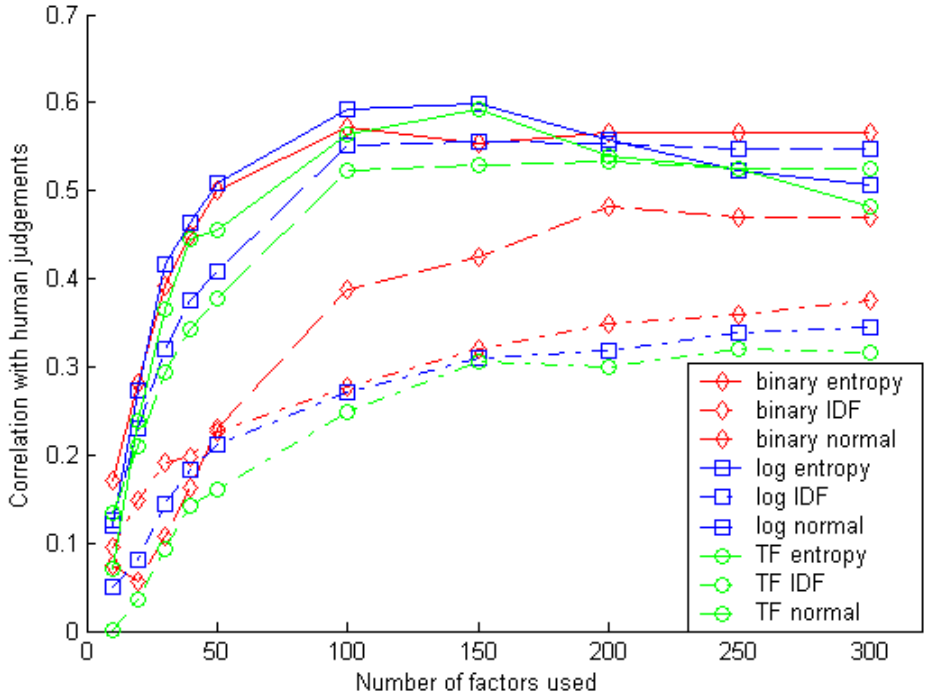

Figure 6: Correlation between stopped, backgrounded LSA and human judgements.

The results from introduction of stopping along with backgrounding are shown in Figure 6. There is an improvement in the correlations over those in Figure 5 for entropy and inverse document frequency global weights and a slight reduction for normal global weighting. The best correlation with human judgements was 0.5988 for log-entropy weighting using 150 factors, closely followed by 0.5935 when using 100 factors and 0.5916 for tf-entropy using 150 factors. Correlations subsequently tail off for these two weighting schemes. The correlations for binary-normal weighting also tail off after a lower peak of 0.4813 at 200 factors. Binary-entropy, log-normal and tf-normal schemes plateaux from 200 factors. Correlations for the idf schemes continue to rise out to 300 factors but are significantly below those of the other six weight combinations. A

comparison of the correlations of idf weighted LSA to human judgements in the presence (Figure 6) and absence (Figure 5) of stopping reveals that stopping significantly improves their performance.

Overall, the two best correlations with human judgements of pairwise document similarity are achieved using log-entropy weighting on stopped and backgrounded text. This is consistent with the literature where log-entropy weighting has performed best in information recall (Dumais, 1991) and text categorisation (Nakov et al., 2001). More controversial are the relative performances of the normal and idf global weighting schemes. The results showed that the use of idf as the global-weight produced correlations with human pairwise judgements that were uniformly worse than those achieved using entropy or normal global-weights in similar situations. In an information recall study (Dumais, 1991) idf weighting outperformed normal weighting. The same is true for most local weighting schemes in a text identification study (Nakov et al., 2001) although this ordering of global weighting function performance did occur for term-frequency local weighting.

Conclusion

LSA can produce pairwise similarities with a higher correlation to human judgments than simple n-gram or bag-of-words vector space methods (Lee, Pincombe, Walsh, 2003). However LSA is more variable and can also have a much lower correlation depending on the choice of weighting functions, factors retained, stopping and backgrounding. The choice of the global weighting function affects the correlations more than any other characteristic. The use of idf global weighting produces correlations with human pairwise judgments that are uniformly worse than those achieved using entropy or normal global-weights in similar situations. Variations in global weights have much more effect on the level of correlation with human pairwise judgments than do variations in local weights. Stopping improves correlations with human for idf and entropy global weights under all circumstances but for normal global weighting correlations only increased when no background documents were used.

None of the combinations of LSA settings produced correlations with human judgements of pairwise document similarity greater than 0.5988 and the worst result was –0.0144. It is evident that care must be taken in the settings used in LSA and that, even when the best settings are used, the results may have only passing similarity to those that would be arrived at by a human.

Acknowledgements

This work has been performed as part of the DSTO contribution to the collaborative research contract "Comparison of Machine and Human Judgements of Relevance". As such the author wishes to acknowledge the work of Michael Lee and Matthew Welsh at Adelaide University that have combined with these results to produce a paper submitted to JASIS briefly covering the correlations with human judgements of a number methods of assessing text similarity. Chris Woodruff, from DSTO, was involved in determining the requirements on the documents contained in the set for which human similarity judgements were obtained. The research assistants and Psychology I students who collected the data and participated as subjects were crucial to both this and other related studies. Phil Radoslovich, Jim Mitkas and Debbie Barnett from the DSTO Business Office and Bruce Ward, Ian Coat and Chris Woodruff from the authors line management have all made this work immeasurably easier through facilitating research cooperation with Michael Lee and his team at Adelaide University. Miro Kraetzl critically assessed the manuscript before it was sent for review.

References

Allan, J., Callan, J., Croft, W.B., Ballesteros, L., Byrd, D., Swan, R. and Xu, J. (1998). INQUERY Does Battle With TREC-6, (pp. 169—207). In Voorhees, E. and Harman, D. (Eds.) *The Sixth Text REtrieval Conference (TREC 6)*. NIST Special Publication 500-240.

Baayen,R.H. (2001). *Word Frequency Distributions*. Kluwer Academic Publishers, P.O. Box 322, 3300 AH Dordrecht, The Netherlands.

Berry, M. W. (1992). Large scale singular value computations. *International Journal of Supercomputer Applications*, 6(1), 13-49.

Berry, M.W., Dumais, S.T. and O'Brien, G.W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review*, **37**(4), 573-595.

Berry, M.W., Drmavc, Z. and Jessup, E.R. (1999). Matrices, vector spaces and information retrieval. *SIAM Review*, **41**(2), 335-362.

Borko, H., and Bernick, M. D. (1963). Automatic document classification. *Journal of the Association for Computing Machinery*, 10:1151–1162.

Carbonell, J., Yang, Y., Frederking, R., Brown, R.D., Geng, Y. and Lee, D. (1997). Translingual Information Retrieval: A Comparative Evaluation. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence*.

Carroll, J.B. (1969). A rationale for an asymptotic lognormal form of word frequency distributions. *Research Bulletin*, Educational Testing Service, Princeton.

Chen, Y. and Leimkuhler, F. (1989). A type-token identity in the Simon-Yule model of text. *Journal of the American Society for Information Science*, **40**, 45—53.

Davis, M.W. and Ogden, W.C. (1998). Free Resources And Advanced Alignment For Cross-Language Text Retrieval (pp.385—402). In Voorhees, E. and Harman, D. (Eds.) *The Sixth Text REtrieval Conference (TREC 6).* NIST Special Publication 500-240.

Deerwester, S.C., S.T. Dumais, T.K. Landauer, G.W. Furnas and R.A. Harshman (1990) "Indexing by Latent Semantic Analysis", *Journal of the American Society of Information Science*, **41**(6), 391—407.

Dumais, S.T. (1990) Enhancing Performance in Latent Semantic Indexing (LSI) Retrieval, TM-ARH-017527 Technical Report, Bellcore.

Dumais, S.T. (1991). Improving the retrieval of information from external sources. *Behaviour Research Methods, Instruments, and Computers*, **23**(2), 229-236.

Dumais, S.T., T.K. Landauer, and M.L. Littman (1996) "Automatic cross-linguistic information retrieval using Latent Semantic Indexing." In *SIGIR'96 - Workshop on Cross-Linguistic Information Retrieval*, pp. 16-23, August 1996.

Gaussier, E., Grefenstette, G., Hull, D.A. and Schulze, B.M. (1998). Xerox TREC-6 Site Report: Cross Language Text Retrieval, (pp. 775-782). In Voorhees, E. and Harman, D. (Eds.) *The Sixth Text REtrieval Conference (TREC 6).* NIST Special Publication 500-240.

Herdan, G. (1964). *Quantative Linguistics*. Buttersworths, London.

Hull, D. and Grefenstette, G. (1996). Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retrieval. (pp. 49--57). In *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland.

Khmaladze, E.V. and Chitashvili, R.J. (1989). Statistical analysis of large number of rare events and related problems. *Transactions of the Tiblisi Mathematical Institute*, **91**, 196—245.

Kintsch, W., Patel, V.L., and Ericsson, K. A. (1999) The role of long-term working memory in text comprehension. *Psychologia* **42**, 186-198.

Kintsch, W.(2000). Metaphor comprehension: A computational theory. *Psyhonomic Bulletin and Review.*, **7**, 257-266.

Kintsch, W.(2001) Predication. *Cognitive Science* **25**, 173-202.

Koll, M. (1979). An approach to concept-based information retrieval. *ACM Special Interest Group on Information Retrieval Forum*, 13:32–50.

Lemaire, B. (1998). Models of High-dimensional Semantic Spaces, In Proceedings of the 4th International Workshop on MultiStrategy Learning (MSL'98), June 1998. (http://citeseer.nj.nec.com/lemaire98models.html)

Foltz, P.W. (1996). Latent semantic analysis for text-based research. *Behaviour Research Methods, Instruments, and Computers*, 28, 197-202.

Hoffman, T. (1999). Probabilistic latent semantic indexing. In *Proc. 22nd Annual Intl. ACM SIGIR Conf. on R&D in Information Retrieval*, pages 50–57.

Laham, D. (1997). Latent Semantic Analysis approaches to categorization. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th annual meeting of the Cognitive Science Society* (p. 979). Mawhwah, NJ: Erlbaum.

Landauer, T. K., and Littman, M. L. (1990). Fully automatic cross-language document retrieval using latent semantic indexing. In *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research* (pp. 31-38). Waterloo, Ontario: UW Centre for the New OED.

Landauer, T.K. and Dumais, S.T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* 104(2) , 211-240.

Landauer, T. K., Laham, D., and Foltz, P. W., (1998). Learning human-like knowledge by Singular Value Decomposition: A progress report. In M. I. Jordan, M. J. Kearns & S. A. Solla (Eds.), *Advances in Neural Information Processing Systems 10,*(pp. 45-51). Cambridge: MIT Press.

Landauer, T. K., Laham, D., & Foltz, P. W., (1998). Learning human-like knowledge by Singular Value Decomposition: A progress report. In M. I. Jordan, M. J. Kearns & S. A. Solla (Eds.), *Advances in Neural Information Processing Systems 10,*(pp. 45-51). Cambridge: MIT Press.

Landauer, T. K., Laham, D., Rehder, B., and Schreiner, M. E., (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In M. G. Shafto and P. Langley (Eds.),*Proceedings of the 19th annual meeting of the Cognitive Science Society* (pp. 412-417). Mawhwah, NJ: Erlbaum.

Littman, M.L., Dumais, S.T. and Landauer, T.K. (1998). Automatic cross-language retrieval using latent semantic indexing. In Grefenstette, G. (Ed.), C*ross-language information retrieval* (chapter 5) Klewer Academic Publishers, Boston.

Mateev, B., Munteanu, E., Sheridan, P., Wechsler, M. and Schäuble P. (1998). ETH TREC-6: Routing, Chinese, Cross-Langauge and Spoken Document Retrieval, (pp. 623—636). In Voorhees, E. and Harman, D. (Eds.) *The Sixth Text REtrieval Conference (TREC 6)*. NIST Special Publication 500-240.

Nakov P. (2000a). Latent Semantic Analysis of Textual Data, (pp. V.3-1-V.3-5). In *Proceedings of CompSysTech'2000*. Sofia, Bulgaria.

Nakov P. (2000b). Getting Better Results with Latent Semantic Indexing, (pp. 156-166). In *Proceedings of the Students Presentations at ESSLLI-2000*. Birmingham, UK.

Nakov P. (2000c). Web Personalization Using Extended Boolean Operations with Latent Semantic Indexing, (pp. 189-198). In *Lecture Notes in Artificial Intelligence - 1904 (Springer). Artificial Intelligence: Methodology, Systems and Applications. 9th International Conference, AIMSA 2000*. Varna, Bulgaria.

Nakov P. (2001a). Latent Semantic Analysis for Bulgarian Literature (pp. 279-284). In *Proceedings of Spring Conference of Bulgarian Mathematicians Union*. Borovetz, Bulgaria.

Nakov P. (2001b). Latent Semantic Analysis for German literature investigation, (pp 834-641). In B. Reusch (Ed.) *7th Fuzzy Days 2001, International Conference on Computational Intelligence*. Dortmund, Germany.

Nakov P., Popova A., Mateev P. (2001). Weight functions impact on LSA performance, (pp. 187-193). In *EuroConference RANLP'2001 (Recent Advances in NLP)*. Tzigov Chark, Bulgaria.

Papadimitriou, C. H., Raghavan, P., Tamaki, H., and Vempala, S. (1998). Latent semantic indexing: A probabilistic analysis. *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*.

Rehder, B., Littman, M.L., Dumais, S.T. and Landauer, T.K. (1998). Automatic 3-Langauge Cross-Language Information Retrieval with Latent Semantic Indexing, (pp. 233—240). In Voorhees, E. and Harman, D. (Eds.) *The Sixth Text REtrieval Conference (TREC 6)*. NIST Special Publication 500-240.

Rehder, B., Schreiner, M. E., Wolfe, M. B., Laham, D., Landauer, T. K., and Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes*, **25**, 337-354.

Schaüble, P. and Sheridan, P. (1998). "Cross-Language Information Retrieval (CLIR) Track Overview" (pp. 31--44). In Voorhees, E. and Harman, D. (Eds.) *The Sixth Text REtrieval Conference (TREC 6)*. NIST Special Publication 500-240.

Sheridan, P. and Ballerini, J.P. (1996). Experiments in Multi-Lingual Information Retrieval using the SPIDER System (pp. 58-65). In *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland.

Sichel, H.S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, **70**, 542—547.

Sichel, H.S. (1986). Word frequency distributions and type-token characteristics. *Mathematical Scientist*, **11**, 45—72.

Sichel, H.S. (1997). Modelling species-abundance frequencies and species-individual functions with the generalized inverse Gauss-Poisson distribution. *South African Statistical Journal*, **31**, 13—37.

Simon, H.A. (1961). Reply to "final note" by Benoit Mandelbrot. *Information and Control*, **4**, 217—223.

Steinhart, D. (2001)Summary Street: an intelligent tutoring system for improving student writing through the use of latent semantic analysis *Unpublished doctoral dissertation, Institute of Cognitive Science, University of Colorado, Boulder*

Terzieva S., Nakov P., Handjieva S. (2001). Investigating the degree of Adequacy of the Relations in the Concept Structure of Students using the Method of Latent Semantic Analysis. In *Proceedings of CompSysTech'2001 - Bulgarian Computer Science Conference*. Sofia, Bulgaria.

Wolfe, M. B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., and Landauer, T. K. (1998). Learning from text: Matching readers and text by Latent Semantic Analysis. *Discourse Processes,* **25**, 309-336.

Zha, H., and Zhang, Z. (1998). On matrices with low-rank-plus-shift structures: Partial SVD and latent semantic indexing. Technical Report CSE-98-002, Department of Computer Science and Engineering, Pennsylvania State University

Appendix A: Processing Related Data

A.1. Raw Documents and Word Counts

1. The national executive of the strife-torn Democrats last night appointed little-known West Australian senator Brian Greig as interim leader – a shock move likely to provoke further conflict between the party's senators and its organisation. In a move to reassert control over the party's seven senators, the national executive last night rejected Aden Ridgeway's bid to become interim leader, in favour of Senator Greig, a supporter of deposed leader Natasha Stott Despoja and an outspoken gay rights activist. (80 words)
2. Cash-strapped financial services group AMP has shelved a $400 million plan to buy shares back from investors and will raise $750 million in fresh capital after profits crashed in the six months to June 30. Chief executive Paul Batchelor said the result was "solid" in what he described as the worst conditions for stock markets in 20 years. AMP's half-year profit sank 25 per cent to $303 million, or 27c a share, as Australia's largest investor and fund manager failed to hit projected 5 per cent earnings growth targets and was battered by falling returns on share markets. (98 words)
3. The United States government has said it wants to see President Robert Mugabe removed from power and that it is working with the Zimbabwean opposition to bring about a change of administration. As scores of white farmers went into hiding to escape a round-up by Zimbabwean police, a senior Bush administration official called Mr Mugabe's rule "illegitimate and irrational" and said that his re-election as president in March was won through fraud. Walter Kansteiner, the assistant secretary of state for African affairs, went on to blame Mr Mugabe's policies for contributing to the threat of famine in Zimbabwe. (98 words)
4. A radical armed Islamist group with ties to Tehran and Baghdad has helped al-Qaida establish an international terrorist training camp in northern Iraq, Kurdish officials say. Intelligence officers in the autonomous Kurdish region of Iraq told the Guardian that the Ansar al-Islam (supporters of Islam) group is harbouring up to 150 al-Qaida members in a string of villages it controls along the Iraq-Iran border. Most of them fled Afghanistan after the US-led offensive, but officials from the Patriotic Union of Kurdistan (PUK), which controls part of north-east Iraq, claim an "abnormal" number of recruits are making their way to the area from Jordan, Syria and Egypt. (106 words)
5. Washington has sharply rebuked Russia over bombings of Georgian villages, warning the raids violated Georgian sovereignty and could worsen tensions between Moscow and Tbilisi. "The United States regrets the loss of life and deplores the violation of Georgia's sovereignty," White House spokesman Ari Fleischer said. Mr Fleischer said US Secretary of State Colin Powell had delivered the same message to his Russian counterpart but that the stern

language did not reflect a sign of souring relations between Moscow and Washington. (80 words)

6. A gay former student of a Melbourne Christian school is taking legal action under equal opportunity legislation, claiming the school discriminated against him because of his sexuality. Tim, 16, alleged a staff member at Hillcrest Christian College in Berwick told him he "had the devil in him", and constant bullying by students prompted the principal to tell him to hide his sexuality. He left the school several weeks ago and is continuing Year 10 by distance education after he said homophobic bullies threw rocks at his head, spat on him, called him names and slashed his belongings. (97 words)
7. Senior members of the Saudi royal family paid at least $560 million to Osama bin Laden's terror group and the Taliban for an agreement his forces would not attack targets in Saudi Arabia, according to court documents. The papers, filed in a $US3000 billion ($5500 billion) lawsuit in the US, allege the deal was made after two secret meetings between Saudi royals and leaders of al-Qa'ida, including bin Laden. The money enabled al-Qa'ida to fund training camps in Afghanistan later attended by the September 11 hijackers. The disclosures will increase tensions between the US and Saudi Arabia. (97 words)
8. Palestinian hired gun Abu Nidal, whose violent death was reported last week from Baghdad, was murdered on the orders of Iraqi President Saddam Hussein after refusing to train al-Qa'ida fighters based in Iraq, reports said yesterday. Iraqi intelligence chief Taher Jalil Habbush said last Wednesday Abu Nidal had shot and killed himself after being discovered living illegally in Baghdad and facing interrogation for anti-Iraqi activities. But Western diplomats believe the radical militant was killed for refusing to reactivate his international terrorist network. (82 words)
9. Hunan province remained on high alert last night as thunderstorms threatened to exacerbate the flood crisis, now entering its fifth day and with 108 already dead and hundreds of thousands evacuated. On the flood frontline at Dongting Lake, the water level peaked at just under 35m on Saturday night, then eased about 3cm during the day under a hot sun, with temperatures reaching 35C. But with the lake still brimming at dangerously high levels, and spilling over the top of its banks in some places, locals were fearful that a thunderstorm and high winds forecast to hit the region last night would damage the dikes. About 1800km of dikes around the lake are all that stand between 10 million people in the surrounding farmland and disaster. (126 words)
10. A U.S.-British air raid in southern Iraq left eight civilians dead and nine wounded, the Iraqi military said Sunday. The military told the official Iraqi News Agency that the warplanes bombed areas in Basra province, 330 miles south of Baghdad. The U.S. Central Command in Florida said coalition aircraft used precision-guided weapons to strike two air defense radar systems near Basra "in response to recent Iraqi hostile acts against coalition aircraft monitoring the Southern No-Fly Zone." (76 words)
11. Iraq and Russia are close to signing a $40 billion economic cooperation plan, Iraq's ambassador said Saturday, a deal that could put Moscow at odds with

the United States as it considers a military attack against Baghdad. The statement by Ambassador Abbas Khalaf came amid indications that Russia, despite its strong support for the post-Sept. 11 antiterrorism coalition, is maintaining or improving ties with Iran and North Korea, which together with Iraq are the countries President Bush has labeled the "axis of evil." (83 words)

12. U.S. intelligence cannot say conclusively that Saddam Hussein has weapons of mass destruction, an information gap that is complicating White House efforts to build support for an attack on Saddam's Iraqi regime. The CIA has advised top administration officials to assume that Iraq has some weapons of mass destruction. But the agency has not given President Bush a "smoking gun," according to U.S. intelligence and administration officials. (67 words)
13. Drug squad detectives have asked the Police Ombudsman to investigate the taskforce that is examining allegations of widespread corruption within the squad. This coincides with the creation of a special unit within the taskforce to track the spending of at least 10 serving and former squad members. The corruption taskforce, codenamed Ceja, will check tax records and financial statements in a bid to establish if any of the suspects have accrued unexplained wealth over the past seven years. But drug squad detectives have countered with their own set of allegations, complaining to the ombudsman that the internal investigation is flawed, biased and over-zealous. (103 words)
14. Queensland senator Andrew Bartlett has launched a last-minutebid to rescue the Australian Democrats from a split that threatens to destroy the party. With nominations for the party leadership to close on Wednesday night, Senator Bartlett met last night with deputy leader Aden Ridgeway to offer him a place on a unity ticket and set up a reform process to begin healing the party's wounds. Party sources said Senator Ridgeway, who turned against former leader Natasha Stott Despoja, is still expected to contest the leadership against one of her two supporters: Senator Bartlett or Brian Greig, installed as interim leader by the party's executive last Thursday. (105 words)
15. Very few women have been appointed to head independent schools, thwarting efforts to show women as good leaders, according to the Victorian Independent Education Union. Although they make up two-thirds of teaching staff, women hold only one-third of principal positions, the union's general secretary, Tony Keenan, said. He believed some women were reluctant to become principals because of the long hours and the nature of the work. But in other cases they were shut out of the top position because of perceptions about their ability to lead and provide discipline. (90 words)
16. The Bush administration has drawn up plans to escalate the war of words against Iraq, with new campaigns to step up pressure on Baghdad and rally world opinion behind the US drive to oust President Saddam Hussein. This week, the State Department will begin mobilising Iraqis from across North America, Europe and the Arab world, training them to appear on talk shows, write opinion articles and give speeches on reasons to end President Saddam's rule. (75 words)

17. Beijing has abruptly withdrawn a new car registration system after drivers demonstrated "an unhealthy fixation" with symbols of Western military and industrial strength - such as FBI and 007. Senior officials have been infuriated by a popular demonstration of interest in American institutions such as the FBI. Particularly galling was one man's choice of TMD, which stands for Theatre Missile Defence, a US-designed missile system that is regularly vilified by Chinese propaganda channels. (73 words)
18. The United Nations World Food Program estimates that up to 14 million people in seven countries - Malawi, Mozambique, Zambia, Angola, Swaziland, Lesotho and Zimbabwe - face death by starvation unless there is a massive international response. In Malawi, as many as 10,000 people may have already died. The signs of malnutrition - swollen stomachs, stick-thin arms, light-coloured hair - are everywhere. (62 words)
19. In Malawi, as in other countries in the region, AIDS is making the effects of the famine much worse. The overall HIV infection rate in Malawi is 19 per cent, but in some areas up to 35 percent of people are infected. A significant proportion of the young adult population is too sick to do any productive work. Malnutrition causes people to succumb to the disease much more quickly than they do in the West, and hunger forces women into prostitution in order to feed their families, making them more vulnerable to contracting the disease. Life expectancy has been reduced to 40 years. (103 words)
20. The United Nations was determined that its showpiece environment summit - the biggest conference the world has ever witnessed - should be staged in Africa. The venue, however, could not be further removed from the grim realities of life in the rest of Africa. Johannesburg's exclusive and formerly whites-only suburb of Sandton is the wealthiest neighbourhood in the continent. Just a few kilometres from Sandton begins the sprawling Alexandra township, where nearly a million people live in squalor. Organisers of the conference, which begins today, seem determined that the two worlds should be kept as far apart as possible. Tight security surrounds Sandton's convention centre and five-star hotels, where world leaders will debate poverty, the environment and sustainable development while enjoying lavish hospitality. (122 words)
21. The Iraqi capital is agog after the violent death of one of the world's most notorious terrorists, but the least of the Palestinian diplomat's worries was the disposal of Abu Nidal's body, which lay on a slab in an undisclosed Baghdad morgue. Abu Nidal's Fatah Revolutionary Council is held responsible for the death or injury of almost 1000 people in 20 countries across Europe and the Middle East in the three decades since he fell out with Yasser Arafat over what Abu Nidal saw as Arafat's willingness to accommodate Israel in the Palestinian struggle.
22. The Federal Government says changes announced today to the work for the dole scheme will benefit participants and taxpayers. Federal Employment Services Minister Mal Brough says that from July 1 those taking part in work

for the dole will be able to perform extra hours to complete their mutual obligation more quickly to access training credits. (61 words)

23. The biowarfare expert under scrutiny in the anthrax attacks declared, "I am not the anthrax killer," and lashed out today against Attorney General John Ashcroft for calling him a "person of interest" in the investigation. For the second time in two weeks, the scientist went before a throng of reporters outside his lawyer's office to profess his innocence and decry the attention from law enforcers that he contends has destroyed his life. (72 words)
24. China said Sunday it issued new regulations controlling the export of missile technology, taking steps to ease U.S. concerns about transferring sensitive equipment to Middle East countries, particularly Iran. However, the new rules apparently do not ban outright the transfer of specific items — something Washington long has urged Beijing to do. (54 words)
25. Nigerian President Olusegun Obasanjo said he will weep if a single mother sentenced to death by stoning for having a child out of wedlock is killed, but added he has faith the court system will overturn her sentence. Obasanjo's comments late Saturday appeared to confirm he would not intervene directly in the case, despite an international outcry. (57 words)
26. An Islamic high court in northern Nigeria rejected an appeal today by a single mother sentenced to be stoned to death for having sex out of wedlock. Clutching her baby daughter, Amina Lawal burst into tears as the judge delivered the ruling. Lawal, 30, was first sentenced in March after giving birth to a daughter more than nine months after divorcing. (61 words)
27. How did 2,300 allegedly unregistered missile warheads come to be stored on a Canadian businessman's anti-terrorism training facility in New Mexico? U.S. and Canadian officials are still trying to figure that out, but one security expert says the mystery is a "chilling" one. David Hudak, 41, was arrested in the United States more than a week ago when, according to court documents, agents searching his property found the warheads stored in crates that were marked “Charge Demolition.” (77 words)
28. The Saudi Interior Ministry on Sunday confirmed it is holding a 21-year-old Saudi man the FBI is seeking for alleged links to the Sept. 11 hijackers. Authorities are interrogating Saud Abdulaziz Saud al-Rasheed “and if it is proven that he was connected to terrorism, he will be referred to the sharia (Islamic) court,” the official Saudi Press Agency quoted an unidentified ministry official as saying. (65 words)
29. Sri Lanka's government will lift a four-year ban on Tamil Tiger rebels on Sept. 6, paving the way for peace talks with the insurgents scheduled for later that month in Thailand, a government minister said Saturday. "We will lift the ban as promised," Minister for Rehabilitation Jayalath Jayawardena told The Associated Press. The lifting of the ban is one of the key rebel conditions for resuming peace negotiations with the government after a hiatus of more than seven years. (79 words)
30. A man accused of making hidden-camera footage up the skirts of women also made child pornography of the worst kind, featuring the rape of children as

young as 6, police said Friday. The latest allegations suggest there's nothing humorous about voyeurs who some may perceive to be making secret videos as a joke, Staff-Insp. Gary Ellis said. "Approximately 20 per cent of voyeurs have also committed sexual assault or rape," Ellis said, reading from a recently released federal government report on criminal voyeurism. (83 words)

31. Police are combing through videotapes trying to spot the gunman dressed in black who shot a 30-year-old man to death at a downtown massage parlour. The victim was hit in the stomach and upper body and died about 3 1/2 hours later in hospital. The woman was not hurt. Police urged business owners to turn over any security-camera videotapes they might have that recorded people on the street at the time. Several such videos are now being reviewed. (78 words)
32. The Federal Government did not regret its actions 12 months on from the Tampa asylum seeker crisis, Small Business Minister Joe Hockey said today. Mr Hockey said the Government was not embarrassed by the Tampa issue, which began on August 27 of last year when the captain of the Norwegian cargo ship rescued more than 400 asylum seekers from an Indonesian ferry north of Christmas Island. (66 words)
33. At least three Democrats are considering splitting from the party while no-one has yet nominated to contest the leadership. Three of the "gang of four" senators who ousted Natasha Stott Despoja from the leadership are considering forming a new "progressive centre" party in the fallout from last week's turmoil. This would leave the Democrats with a rump of three or four members. West Australian Senator Andrew Murray said yesterday unless the Democrats left wing gave ground the party would split. (80 words)
34. A young humpback whale remained tangled in a shark net off the Gold Coast yesterday, despite valiant efforts by marine rescuers. With its head snared by the net and an anchor rope wrapped around its tail, the stricken whale was still swimming but hopes for its survival were fading. A second rescue attempt was planned for dawn today after rescuers braved heavy seas, strong wind and driving rain to try to free the whale. (74 words)
35. Prince William has told friends his mother was right all along to suspect her former protection officer of spying on her and he doesn't want any detective intruding on his own privacy. William and Prince Harry are so devastated by the treachery of Ken Wharfe, whom they looked on as a surrogate father, they are now refusing to talk to their own detectives. (63 words)
36. The spectre of Osama bin Laden rose again today, urging Afghans to launch a new Jihad, or holy war, and predicting the fall of the United States, in a hand-written "letter" posted on an Islamic website. There was no hard proof that the scruffy missive was genuine, but IslamOnline.net said it had been received by their correspondent in Jalalabad, eastern Afghanistan, from an Afghan source who asked to remain anonymous. The source claimed it was the "most recent letter" from the world's most wanted man. (85 words)
37. The Johannesburg Earth Summit is set to get under way with the promise that leaders will take action on the environment, debt and poverty. South African

President Thabo Mbeki, speaking at the opening ceremony, said: "Out of Johannesburg and out of Africa must emerge something that takes the world forward." But the absence of US President George W Bush was threatening to overshadow the summit. (65 words)

38. Robert Mugabe strengthened his hold on the Zimbabwean government yesterday by retaining the most combative hardliner ministers in a cabinet shuffle which offered little hope of a moderation of the land seizures and other policies that have kept Zimbabwe in crisis and brought international condemnation. (51 words)
39. They dress in black and disguise their identities with bandannas and sunglasses. Their logo is an image of the Southern Cross constellation, superimposed with a pair of crossed boomerangs, which resembles a swastika. The Blackshirts are former husbands aggrieved by their treatment at the hands of their ex-wives and the courts, who regard themselves as the vanguard of a "men's rights" movement in Australia and say that their actions will be remembered as marking a turning-point in history. (78 words)
40. The real level of world inequality and environmental degradation may be far worse than official estimates, according to a leaked document prepared for the world's richest countries and seen by the Guardian. It includes new estimates that the world lost almost 10% of its forests in the past 10 years; that carbon dioxide emissions leading to global warming are expected to rise by 33% in rich countries and 100% in the rest of the world in the next 18 years; and that more than 30% more fresh water will be needed by 2020. (93 words)
41. Researchers conducting the most elaborate wild goose chase in history are digesting the news that a bird they have tracked for over 4,500 miles is about to be cooked. Kerry, an Irish light-bellied Brent goose, was one of six birds tagged in Northern Ireland in May by researchers monitoring the species' remarkable migration. Last week, however, he was found dead in an Inuit hunter's freezer in Canada, still wearing his £3,000 satelite tracking device. Kerry was discovered by researchers on the remote Cornwallis Island. They picked up the signal and decided to try to find him. (96 words)
42. Russia defended itself against U.S. criticism of its economic ties with countries like Iraq, saying attempts to mix business and ideology were misguided. “Mixing ideology with economic ties, which was characteristic of the Cold War that Russia and the United States worked to end, is a thing of the past,” Russian Foreign Ministry spokesman Boris Malakhov said Saturday, reacting to U.S. Defense Secretary Donald Rumsfeld's statement that Moscow's economic relationships with such countries sends a negative signal. (77 words)
43. Pope John Paul II urged delegates at a major U.N. summit on sustainable growth on Sunday to pursue development that protects the environment and social justice. In comments to tourists and the faithful at his summer residence southeast of Rome, the pope said God had put humans on Earth to be his administrators of the land, “to cultivate it and take care of it.” “In a world ever more interdependent, peace, justice and the safekeeping of creation cannot but

be the fruit of a joint commitment of all in pursuing the common good," John Paul said. (96 words)

44. The Russian defense minister said residents shouldn't feel threatened by the growing number of Chinese workers seeking employment in the country's sparsely populated Far Eastern and Siberian regions. There are no exact figures for the number of Chinese working in Russia, but estimates range from 200,000 to as many as 5 million. Most are in the Russian Far East, where they arrive with legitimate work visas to do seasonal work on Russia's low-tech, labor-intensive farms. (75 words)
45. Australian spies listened to conversations between Norway's ambassador and its foreign office during the Tampa crisis, a soon to be published book will reveal. Phone calls were tapped by the Defence Signals Directorate when Norwegian ambassador Ove Thorsheim visited the freighter during the stand-off. A book, Tampa, to be published in Norway in October, recounts the events which triggered Australia's Pacific Solution and transformed Tampa Captain Arne Rinnan into a homeland hero. (72 words)
46. Batasuna, a political party that campaigns for an independent Basque state, faces a double blow today: the Spanish parliament is expected to vote overwhelmingly in favour of banning the radical group, while a senior investigative judge is poised to suspend Batasuna's activities on the grounds that they benefit Eta, the outlawed Basque separatist group. (56 words)
47. The river Elbe surged to an all-time record high Friday, flooding more districts of the historic city of Dresden as authorities scrambled to evacuate tens of thousands of residents in the worst flooding to hit central Europe in memory. In the Czech Republic, authorities were counting the cost of the massive flooding as people returned to the homes and the Vlava river receded, revealing the full extent of the damage to lives and landmarks. (74 words)
48. The European Parliament is spoiling for a fight with Israel. It has voted to review the EU's diplomatic links with the Jewish state, to impose an arms embargo and to threaten wider trade sanctions. Many MEPs want to go further and dispatch a European military force to the region in order to "protect the Palestinian people". (58 words)
49. Australia's Commonwealth Bank on Wednesday said it plans to cut about 1,000 jobs even as it reported its profit rose 11 percent last fiscal year. Workers reacted angrily to the planned cuts, which Australia's second largest bank said were designed to control costs. The cuts will take effect this financial year. The bank reported net profit of 2.66 billion Australian dollars ($1.4 billion) in the year to June 30, up from 2.4 billion Australian dollars in the previous year. (79 words)
50. Labor needed to distinguish itself from the Government on the issue of asylum seekers, Greens leader Bob Brown has said. His Senate colleague Kerry Nettle intends to move a motion today - on the first anniversary of the Tampa crisis - condemning the Government over its refugee policy and calling for an end to mandatory detention. "We Greens want to bring the Government to book over

its serial breach of international obligations as far as asylum seekers in this country are concerned," Senator Brown said today. (86 words)

A.2. Stop Words

a about all also although am an and another any anybody anyhow anyone anything anywhere are as at b be become been being but by c can cannot could d did do does doing done e each eg either else et etc every ex f for from g h had has have having he hence her hers herself high him himself his how however i ie if in inc indeed is it its j k l ltd m many may me might more mr mrs ms must my myself n no nor not o of oh or otherwise ought our ours ourselves p per put q r re s self selves shall she should sl so some somehow such sup t than that the their theirs them themselves then there therefore these they this those though thus to u us v very via viz vs w was we were what whatever when whence whenever where whereafter whereas whereby wherein whereupon wherever whether which whichever while whither who whoever whole whom whose why will with within without would x y yes you your yours yourself yourselves z

Appendix B: Raw Result Tables

B.1. LSA with 10, 20, 30, 40 and 50 factors for unstopped, lowercase, alphabetic data with word length from 2 to 20.

factors	10	20	30	40	50
binary-entropy	0.2140	0.3450	0.3350	0.3484	0.4713
binary-idf	0.1162	0.1390	0.1311	0.1604	0.1811
binary-normal	0.1762	0.3439	0.3254	0.3350	0.4535
log-entropy	0.1644	0.3639	0.3638	0.3753	0.4942
log-idf	0.0885	0.0980	0.1055	0.1116	0.1111
log-normal	0.1864	0.3433	0.3182	0.3423	0.4709
tf-entropy	0.1559	0.3174	0.3661	0.3839	0.4948
tf-idf	0.0850	0.0813	0.0794	0.0824	0.0833
tf-normal	0.1895	0.3408	0.3191	0.3332	0.4646

B.2. LSA with 10, 20, 30, 40 and 50 factors for stopped, lowercase, alphabetic data with word length from 2 to 20.

factors	10	20	30	40	50
binary-entropy	0.2113	0.3792	0.3628	0.3950	0.5083
binary-idf	0.1679	0.3100	0.3509	0.3867	0.4385
binary-normal	0.1935	0.3720	0.3399	0.3557	0.4631
log-entropy	0.1238	0.3468	0.3953	0.4266	0.5324
log-idf	0.1293	0.2489	0.3164	0.3808	0.4198
log-normal	0.1987	0.3688	0.3418	0.3893	0.4960
tf-entropy	0.0843	0.3192	0.3792	0.4300	0.5347
tf-idf	0.0473	0.2037	0.2675	0.3597	0.3951
tf-normal	0.1931	0.3548	0.3300	0.3780	0.4874

B.3. LSA with 10, 20, 30, 40, 50, 100, 150, 200, 250 and 300 factors for unstopped, lowercase, alphabetic data with word length from 2 to 20.

factors	10	20	30	40	50	100	150	200	250	300
binary-entropy	0.1597	0.2486	0.3380	0.3853	0.4330	0.5499	0.5312	0.5646	0.5076	0.4850
binary-idf	-0.0144	0.0438	0.1121	0.0972	0.0991	0.1041	0.1005	0.1031	0.0973	0.1124
binary-normal	0.0804	0.0519	0.1284	0.1670	0.2223	0.3938	0.4133	0.4860	0.4792	0.4792
log-entropy	0.1102	0.2425	0.3324	0.3786	0.4347	0.5476	0.5740	0.5577	0.5110	0.4956
log-idf	0.0812	0.0871	0.1001	0.1036	0.1053	0.1095	0.1091	0.1047	0.1032	0.1055
log-normal	0.1046	0.2200	0.2792	0.3312	0.3688	0.5320	0.5623	0.5535	0.5535	0.5535
tf-entropy	0.0580	0.2076	0.3083	0.3500	0.3717	0.4889	0.5743	0.5185	0.5234	0.5234
tf-idf	0.0948	0.0905	0.0919	0.0927	0.0931	0.0958	0.0953	0.0919	0.0905	0.0900
tf-normal	0.1146	0.2151	0.2663	0.3173	0.3600	0.5152	0.5177	0.5400	0.5400	0.5400

B.4. LSA with 10, 20, 30, 40, 50, 100, 150, 200, 250 and 300 factors for stopped, lowercase, alphabetic data with word length from 2 to 20.

factors	10	20	30	40	50	100	150	200	250	300
binary-entropy	0.1714	0.2811	0.3917	0.4497	0.5000	0.5713	0.5544	0.5661	0.5661	0.5661
binary-idf	0.0947	0.1479	0.1910	0.1972	0.2255	0.2764	0.3199	0.3490	0.3601	0.3754
binary-normal	0.0747	0.0563	0.1081	0.1636	0.2298	0.3884	0.4237	0.4813	0.4704	0.4704
log-entropy	0.1202	0.2731	0.4174	0.4628	0.5083	0.5935	0.5988	0.5578	0.5223	0.5058
log-idf	0.0492	0.0810	0.1452	0.1836	0.2113	0.2712	0.3095	0.3186	0.3382	0.3448
log-normal	0.1264	0.2309	0.3213	0.3760	0.4085	0.5507	0.5555	0.5531	0.5469	0.5469
tf-entropy	0.0711	0.2393	0.3659	0.4455	0.4553	0.5643	0.5916	0.5396	0.5253	0.4827
tf-idf	0.0008	0.0349	0.0924	0.1429	0.1604	0.2497	0.3050	0.2993	0.3193	0.3155
tf-normal	0.1342	0.2088	0.2935	0.3435	0.3779	0.5236	0.5285	0.5323	0.5246	0.5246

Page classification: UNCLASSIFIED

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION
DOCUMENT CONTROL DATA

1. PRIVACY MARKING/CAVEAT (OF DOCUMENT)

2. TITLE

Comparison of Human and LSA Judgements of Pairwise Document Similarities for a News Corpus

3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION)

Document	(U)
Title	(U)
Abstract	(U)

4. AUTHOR(S)

Brandon Pincombe

5. CORPORATE AUTHOR

Information Sciences Laboratory
PO Box 1500
Edinburgh South Australia 5111 Australia

6a. DSTO NUMBER	6b. AR NUMBER	6c. TYPE OF REPORT	7. DOCUMENT DATE
DSTO-RR-0278	AR-013-177	Research Report	September 2004

8. FILE NUMBER	9. TASK NUMBER	10. TASK SPONSOR	11. NO. OF PAGES	12. NO. OF REFERENCES
E 8709/9/21	INT 03/082	POLCOM	36	51

13. URL on the World Wide Web

http://www.dsto.defence.gov.au/corporate/reports/DSTO-RR-0278.pdf

14. RELEASE AUTHORITY

Chief, Intelligence, Surveillance and Reconnaissance Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for public release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CITATION IN OTHER DOCUMENTS Yes

18. DEFTEST DESCRIPTORS

Semantics
Human information processing
Knowledge acquisition
Word recognition
Automatic indexing

19. ABSTRACT

Pairwise similarity judgement correlations between humans and Latent Semantic Analysis (LSA) were explored on a set of 50 news documents. LSA is a modern and commonly used technique for automatic determination of document similarity. LSA users must choose local and global weighting schemes, the number of factors to be retained, stop word lists and whether to background. Global weighting schemes had more effect than local weighting schemes. Use of a stop word list almost always improved performance. Introduction of a background set of similar documents increased larger correlations and reduced smaller ones. The correlations ranged between approximately 0 and 0.6 depending on the LSA settings indicating the importance of correct settings. The low maximum correlation indicates that information presentation schemes based on LSA may often be at variance with visualisations based on human decisions even using the best settings for a data set.

Page classification: UNCLASSIFIED