Cross-lingual latent semantic analysis

W. Cox[1] B. Pincombe[2]

(Received 31 August 2006; revised 10 November 2008)

Abstract

Cross-lingual information retrieval is a difficult task typically involving query translation into multiple languages followed by monolingual retrieval in each language. Latent Semantic Analysis allows cross-lingual retrieval without translating queries by working from an already existing corpus of translations. Thus, collecting such a corpus obviates the need to construct complicated translation tools, making this technique particularly applicable to querying less commercially appealing languages. First, we extend work on retrieval from an English-French corpora split into training and test sets to examine the effects of training on a corpus from a completely different. Success is measured by the proportion of direct translations correctly considered most similar by Latent Semantic Analysis. Secondly, an English only similarity task from the literature is also extended to train on a different corpus to the one being tested on. Here the degradation in

http://anziamj.austms.org.au/ojs/index.php/ANZIAMJ/article/view/98 gives this article, © Austral. Mathematical Soc. 2008. Published December 4, 2008. ISSN 1446-8735. (Print two pages per sheet of paper.)

performance is measured through examining the variation in the correlations between the inter-document similarity judgements calculated by Latent Semantic Analysis and an experimentally derived baseline of human judgements of inter-document similarity. Higher order indexing schemes discarding uncommon terms, sparse matrix representations and the removal of factors with very low eigenvalues are used to enhance efficiency. Performance degradation from exogenous training is shown in both cases. The best results occur using stopping, log-entropy weighting and over 500 factors.

Contents

1 Introduction

Latent Semantic Analysis (LSA) is an automated technique for comparing the similarity of documents. It has been used in document visualisation tools [1], library retrieval tools [1], SPAM filters [7] and to automatically grade student essays [6]. A relatively new application of LSA involves assessing cross-lingual document similarities [3, 5]. This is known as Cross-Lingual LSA (CL-LSA)

and caters for situations requiring searching in one language and retrieval in another. This is particularly important for languages that do not have quality machine translation tools available for them such as those from lesser developed regions. The United Nations is one example of a body that could use such a capacity. With CL-LSA we only need to translate the documents retrieved, rather than all the queries going into the system. Since there are few documents to be translated it is more feasible to use skilled linguists.

A high level of accurate pairing of documents and their translations has been demonstrated using an English-French corpora split into training and test sets [3]. However, one of the problems of applying CL-LSA in the real world is that there are likely to be many situations where the document set being queried is not particularly closely related to the cross language document set used in training. This is particularly the case in languages of small and economically less developed groups where there are limited sets of well translated documents. Therefore, training with one half of a corpus and withholding the other half for testing gives an upper bound on the performance. It would be nice to have some idea of the level of degradation of performance introduced by training with a corpus from one subject area and testing on a corpus from another. This is explored for a single case in this article and the level of degradation in performance is found to be in the order of 20%.

We assess the performance of mono- and cross-lingual LSA for testing of cross-corpus document similarities and identify the parameters which optimise the performance of LSA in some of these circumstances. Training and testing are performed on different corpora both for CL-LSA and for LSA. CL-LSA uses the English-French retrieval task previously performed in the literature [3] on Canadian Hansard except that a corpus of Amnesty International press releases is used for training. Performance is measured by the proportion of direct translations considered most similar. In the mono-lingual case the similarity judgements produced by LSA were compared to a human baseline gathered elsewhere [10]. These comparisons were performed for LSA trained

on documents within the test set, from without the test set but within the ABC Newsmail corpus it was drawn from and from the English part of the Canadian Hansard.

The assessment techniques used for LSA and CL-LSA differed. For LSA correlations with human judgements of document similarity were used, whereas for CL-LSA, ability to retrieve the exact translation as the most similar document was looked at. There was a massive degradation in the ability of LSA to predict human judgements of document similarity when training was performed outside the corpus used for testing. The use of stopping, log-entropy weighting and over 500 factors were found to be the optimal parameters which is consistent with the literature [3, 4, 9, 10]. The Matlab code written to do this never reached the levels of performance on this task reported in the literature [10] for Perl, shell script and C code using SVDPACK. This is an important point as the CL-LSA code was based on the Matlab version of LSA. Considering the performance of CL-LSA, there was a degradation of approximately 20% in the level of matching of documents with their direct translations when a different corpus was used for training.

2 Latent semantic analysis

LSA involves constructing a term-document matrix for a large collection of documents. This matrix gives the number of occurrences of each term (which are essentially the same as words) within each document. Singular value decomposition is then used to construct the semantic space for the corpus. LSA assumes that there is some amount of noise present in all natural language as a result of different authors using different words to express the same concept. It attempts to remove this noise and represent the underlying concepts within documents via re-multiplication of the decomposition matrices using a reduced number of factors.

Perhaps the greatest benefit of LSA is its ability to overcome the funda-

mental problems of synonymy, polysemy and inflexion which are inherent in natural language processing. Synonymy means that many different words have similar meaning. For example, if we searched for the word "large", term-matching techniques would not retrieve relevant documents containing the words "big", "huge" or "massive". However, LSA recognises that these words all refer to the same concept and hence would retrieve all of these documents.

A polysemous word is one which has several meanings depending on the context. For example, the word "chip" could be referring to fish and chips, a computer chip, a gambling chip or a chip of wood, depending on the context. If we wanted information about computer chips and simply searched for the word "chip", term-matching techniques would retrieve irrelevant documents related to the other meanings of "chip".

Inflexion is the process of adding affixes to or changing the base form of a word [4]. The words "doing", "did", "doer", and "do" are all related to "done", but would not be retrieved by term-matching techniques upon searching for this term without the use of some form of stemming. While stemming techniques vary in accuracy, the majority are based on rules of grammar and are therefore not cross lingual. Only co-occurrence based stemmers can be applied over multiple languages.

Although LSA has many advantages, there are also some limitations. Firstly, LSA is hampered by the processing power and memory capacity of computers. Empirical results suggest that the larger the size of the training set, the better the performance of LSA. However, constructing a term-document matrix for a relatively small set of 2000 documents can take up to an hour and use 300 Mb of memory. It is hoped that due to the continual increase in the speed and memory of computers this will not be such a problem in the future.

Secondly, LSA is a "bag of words" technique which means that it makes no use of word order. Hence, we expect it to miss some of the concepts within

documents. At the same time though, this highlights the amazing power of LSA that it can be exposed to nothing but a set of words and manage to infer deep relations between the structure and meanings of words and documents [5].

Thirdly, LSA is limited by the size of the text corpora used. Apart from taking a long time to process, sufficiently large text corpora simply may not be readily available for specific fields.

As mentioned above, cross-lingual LSA allows us to assess similarities between documents in multiple different languages. It involves training on a set of parallel documents in two or more different languages. Queries can then be given in a native language and documents retrieved in that language as well as other languages. No translation is required for the document retrieval since CL-LSA produces a model of the conceptual content of documents which is language independent [5].

Past tests of CL-LSA have produced promising results. Applying CL-LSA to French-English Canadian Hansard documents over 98% of five-word queries were able to retrieve their cross-language mate [3]. Encouraging results have arisen from applying CL-LSA to Greek-English versions of the Gospel [3] and English-Japanese extracts of scientific articles [5]. We test the performance of CL-LSA when the pseudo-documents are not from the original corpus.

3 Method

The same general process can be applied to obtain mono-lingual and cross-lingual document similarity judgements using LSA.

Firstly, we need to obtain a background corpus of documents. The performance of LSA depends immensely on the diversity and applicability of

Amnesty International deplores the decision of the Interim Government of Iraq to reimpose the death penalty and believes that it will do nothing to restore security for the people of Iraq.	Amnesty International déplore la décision du gouvernement intérimaire irakien de rétablir la peine de mort et estime que cela ne permettra pas de rétablir la sécurité dans le pays.

FIGURE 1: Parallel English and French passages from an Amnesty International press release.

this background set to future queries. The experiments presented here used ABC newsmail reports along with English and French Canadian Hansard proceedings and also English and French Amnesty International press releases. An example of parallel passages from an Amnesty International document is given in Figure 1.

Second, we conduct some pre-processing of the documents in our set. This involves formatting the documents and removing stop words and is achieved using a text processor such as Perl. Stop words are small, commonly used words, such as "a", "and", "is", "to" and "the", which do not contribute much to document meaning. It has been demonstrated that removal of these words improves the performance of LSA [8]. For the cross-lingual case, we also need to match up parallel documents.

The third step in assessing document similarities using LSA is the construction of the term-document matrix. This gives the number of occurrences of each word within each document from the corpus. The construction of this matrix and the following steps below were accomplished using Matlab. Figure 2 demonstrates the initial step of computing term occurrence counts for a small corpus of three documents. First the text at the top of Figure 2 is stopped, thus removing the non-italicised words from the conceptual word document matrix at the lower left and from the word document matrix at the lower right.

Doc 1 "The *boy* went to the *shop* but the *shop* was *closed*"

Doc 2 "The *girl* went to *school* to use a *computer*"

Doc 3 "A *computer* can be *purchased* from the *shop*"

	Doc 1	Doc 2	Doc 3
boy	x		
shop	xx		x
closed	x		
girl		x	
school		x	
computer		x	x
purchased			x

$$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

FIGURE 2: Three sample documents (top); the word-document matrix after stop words are removed (lower left); and the numerical representation (lower right).

Weighting of this term-document matrix further improves the performance of LSA. Firstly, global weighting of each term across all documents takes place and then local weighting of each term within each document. The resulting total weight w_{ij} of term i in document j is

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

where $L(i,j)$ is the local weight of term i within document j and $G(i)$ is the global weight of term i across all documents [8]. We used entropy global weighting combined with term-frequency, logarithmic and binary local weighting as these schemes have been identified as giving the best performance [8]. Results were then compared against those obtained using no weighting of the term-document matrix.

Entropy global weighting is defined as

$$G(i) = 1 - \frac{H(d \mid i)}{H(d)} \tag{2}$$

where $H(d \mid i) = -\sum_{k=1}^{J} p(i,k) \log_2 p(i,k)$ is the entropy of the conditional distribution given i and $H(d) = \log_2 J$ is the entropy of the document distribution [8]. Logarithmic local weighting is

$$L(i,j) = \log_2 \left(\mathrm{tf}(i,j) + 1\right) \tag{3}$$

where

$$\mathrm{tf}(i,j) = \frac{c(i,j)}{\sum_{k=1}^{I} c(k,j)} \tag{4}$$

is the term frequency of the ith term in the jth document, $c(i,j)$ and $c(k,j)$ are the number of appearances of the ith and kth terms in the jth document and I is the total number of terms [8]. Term-frequency and binary local weighting resulted in lower performance than logarithmic local weighting and hence are not presented here. The equations for these schemes have been given by Pincombe [8].

Singular value decomposition is the core mathematical technique which LSA is based upon. It involves decomposing our term-document matrix into the product of three other matrices:

$$\mathbf{X} = \mathbf{T} \times \mathbf{S} \times \mathbf{D}^{\mathrm{T}} \tag{5}$$

where $\mathbf{T}$ is the term decomposition matrix, $\mathbf{S}$ is the matrix of singular values and $\mathbf{D}$ is the document decomposition matrix [2]. $\mathbf{X}$ is the original weighted term-document matrix. Performing the singular value decomposition constitutes a large proportion of the total execution time, but we should only have to perform it once. The following step of assessing similarities between test documents is very quick. We hoped that our cross-corpus document similarity testing would allow us to assess whether or not it is necessary to re-perform the lengthy decomposition process when new test documents are obtained.

Test documents (or pseudo-documents) are documents which we wish to assess similarities between. They are additional to the background set. The next step in the LSA process involves folding these test documents into our current representation for the background set. We do this by identifying the terms from the background set which occur in the set of test documents. A pseudo term-document matrix containing only these terms is then produced. In effect, we are placing pseudo-documents at the centroid (or weighted average) of their constituent terms in the latent semantic space constructed previously.

Following the construction of the pseudo term-document matrix ($\mathbf{X}_q$) we can produce a reduced dimensionality approximation of this matrix without actually performing a singular value decomposition [2]. Instead, we just compute a pseudo document decomposition matrix ($\mathbf{D}_q$) using

$$\mathbf{D}_q = \mathbf{X}_q^{\mathrm{T}} \times \mathbf{T} \times \mathbf{S}^{-1} \tag{6}$$

and then use this matrix along with the original decomposition matrices $\mathbf{S}$ and $\mathbf{T}$ to calculate the approximation

$$\hat{\mathbf{X}}_q = \mathbf{T} \times \mathbf{S} \times \mathbf{D}_q^{\mathrm{T}} \tag{7}$$

This matrix, $\hat{\mathbf{X}}_q$, is more valuable to us than the original pseudo term-document matrix, $\mathbf{X}_q$, as the lower dimensions more accurately reveal the relationships between terms and documents.

The final step in the LSA process is to compute a matrix of pair-wise document similarities. We do this simply by multiplying $\hat{\mathbf{X}}_q$ with its transpose, which calculates the dot product of each document vector with every other document vector. Normalisation is then performed so that all document similarity values are in $[-1, 1]$.

To assess the performance of mono-lingual LSA, the correlation between the LSA computed similarity matrix and a matrix of human judgements of document similarities is calculated. A high correlation indicates that LSA has performed well. Performance in the cross-lingual case is measured by computing the percentage of documents which correctly identify their cross-language mate as having the highest similarity to themselves.

Refinements were made to enhance efficiency. The number of index terms was decreased by discarding those terms which only appear in very few documents (we experimented with discarding terms which appeared in less than n documents, for $n = 2, 3, 4, \ldots$ —we give these indexing schemes the name "*ndoc*"). Factors with very low eigenvalues (where there is a drop in order of magnitude of consecutive eigenvalues greater than 10^6) were removed since these do not contribute much to the re-construction of the term-document matrix anyway. Also, sparse matrix representations were used to save storage space. This was implemented after discovering that typically less than 10% of the entries in the term-document matrix were non-zero.

4 Results and discussion

Both LSA and CL-LSA suffered a reduction in performance when trained on documents from one corpus and tested on documents from another. It is

difficult to compare the level of reduction in performance as similarity to human judgements was tested for LSA and retrieval was tested for CL-LSA.

A large variety of parameters was used in LSA in a search for optimal performance but only the best outcome is displayed in Figure 3. As could be expected [8, 10], log-entropy weighting (logarithmic local weighting and entropy global weighting) and an indexing scheme which retained all terms produced the best results. As the human judgements of document similarity [10] were only performed on ABC Newsmail documents we were locked into using these as the test set. Three types of training set were used, all of them containing 2,482 documents.

The best results occurred when 2,432 documents were chosen at random from the ABC Newsmail data set and augmented by the 50 ABC Newsmail documents used in the document similarity study. This process was performed 50 times and the average results are shown in the top line in Figure 3. The best correlation with human judges was approximately 0.47 which is below the correlation of approximately 0.60 found using different code on the same dataset [8, 10] with a smaller number of background documents. While there may be minor problems with the code, remember that the use of 364 documents in the literature gives a greater weight to the 50 documents that the test is being performed upon. As the 314 non-test documents in the literature were drawn from the stories in the same two month period as the test documents it is more likely they would share themes with the test documents than would 2,432 documents drawn from a four year period. Note that people only managed a correlation of 0.62 with the average person as there was considerable disagreement on how related documents were to each other.

The second best result occurred when all 2,482 documents were chosen at random from the set of ABC Newsmail documents excluding the 50 documents used in the test set. Again the line in Figure 3 represents the average of 50 runs. Maximal performance was just below 0.43 and the drop off in performance from the loss of the 50 documents used in testing lends weight to the argument that the difference between previously published results [8, 10]

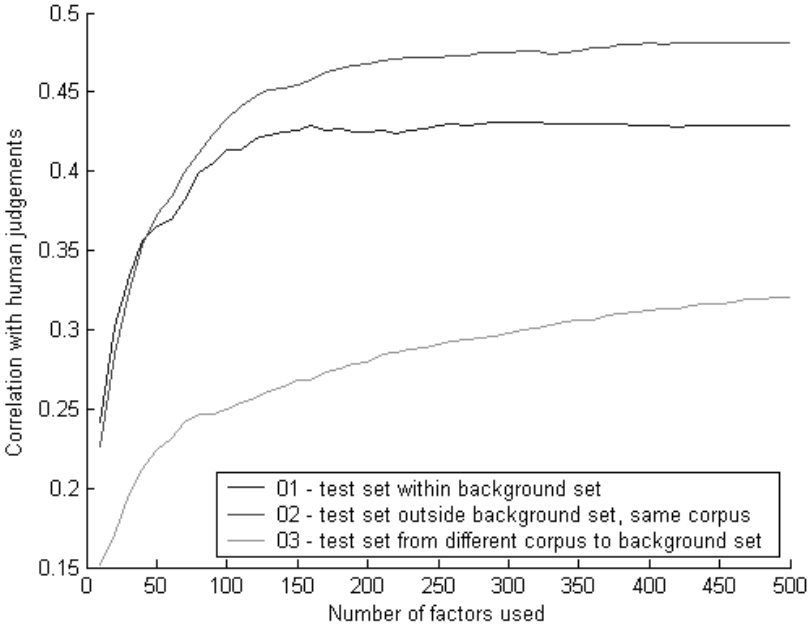

FIGURE 3: Correlation of LSA-computed similarities with human baseline using a training set consisting of ABC newsmail documents, log-entropy weighting and an indexing scheme which retains all terms.

and those in the first case discussed above are mostly due to the presence of more documents that are unrelated to the test set than is the case in the literature. As it has been identified that backgrounding of LSA with documents from outside the test set improves performance over use of the test set only [8, 10] it raises the interesting conjecture that at some point backgrounding becomes baneful rather than beneficial.

The lowest line in Figure 3 is for the case where the 2,482 documents were chosen from the English members of the Canadian Hansard. The results are poor. Although they are still increasing at 500 factors they only just

exceeded 0.3 correlation with human judges. This further degradation in performance indicates the problems of using a backgrounding or training set that is less related to the documents being tested. Many of the low frequency terms important for indicating meaning and judging the nuances of similarity that are present in the 50 document set are simply not present in the Canadian Hansard.

The less related the training corpus is to the test set the poorer the results that are achieved.

Matrix processing constraints for the term-document matrices in CL-LSA necessitated an indexing scheme dumping less common terms (in this case all terms were words). This was necessary because the use of two languages doubled the number of terms derived from the document set. The "*4doc*" indexing scheme was used whereby only those terms appearing in four or more documents were retained. Log-entropy weighting was also used after being identified as the optimal weighting scheme from the mono-lingual LSA analysis.

The bar graph in Figure 4 presents the proportion of documents which correctly identified their cross-language mate as having the highest similarity to themselves with an error bar of two standard deviations. Three different scenarios were investigated involving training and testing within the Canadian Hansard set and within the Amnesty International set as well as training on documents from the Amnesty International set and testing on the Canadian Hansard set. In all cases the training was performed on documents formed from concatenating the English and French texts of each given document. The testing was done by sequentially using all of the remaining English documents as queries against a set made up of all the remaining documents, both English and French, apart from the query document. A success was recorded if the most similar document returned was the French translation of the English query. The two sets each contained $1,241$ pairs for a total of $2,482$ pairs of documents.

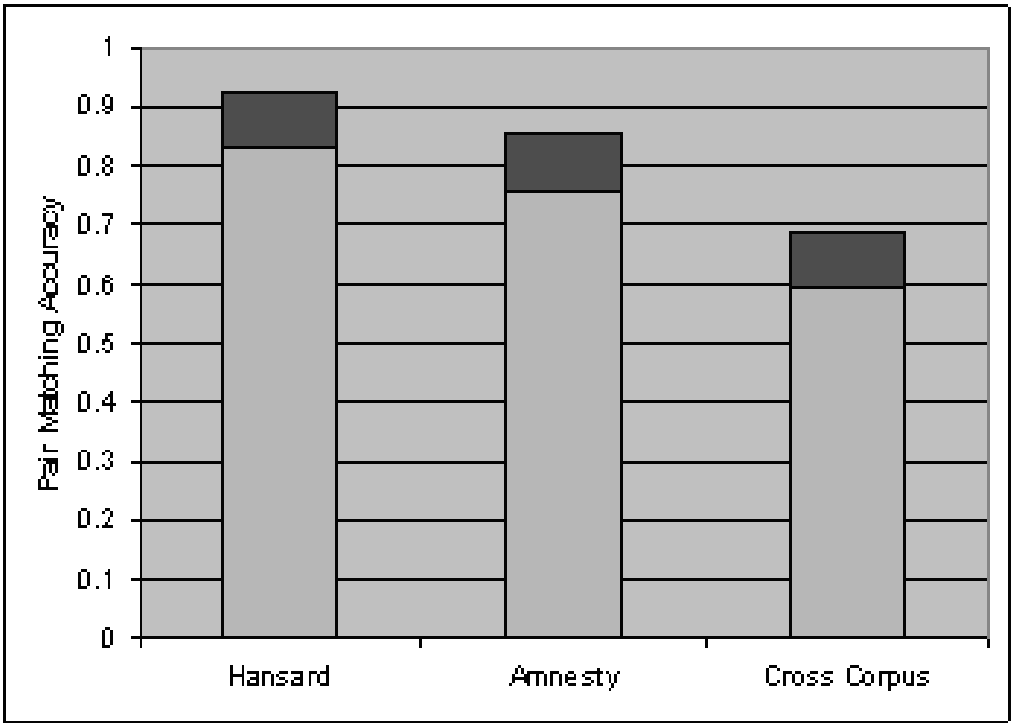

FIGURE 4: Pair matching accuracy for within corpus and cross-corpus testing of English and French documents using CL-LSA (with error bars of two standard deviations)

This process was performed thirty times on the Canadian Hansard documents to obtain the bar graph shown in Figure 4. In each case 1,241 documents were randomly allocated to the training and test sets. Those being trained on were paired with their partner and LSA was performed using the specifications above. The performance was not as good as that reported previously [3]. This was not because of error because the two standard deviation error bar on the mean performance does not overlap with the single figure reported previously. It is most likely due to our use of a higher order indexing scheme (which only retained terms appearing in four or more documents) rather than taking into account the terms appearing in two or more documents as done by Dumais et al. [3]. Some part of it could relate to the minor accuracy problems identified in the LSA section and due to the specific implementation in Matlab rather than the Perl, shell script and C wrapper for SVDPACK used elsewhere [2, 3, 4, 5, 6, 8, 10]. These pairings were correct approximately 87% of the time.

The second bar in Figure 4 shows the results for thirty iterations of the same process applied to 2,482 Amnesty International press releases. In this case the accuracy was lower. This could be related to the lower quality of the data which was disaggregated in transmission and needed to have its pairings reconstructed. The pairings were incorrect in $1.5 \pm 1\%$ of the documents according to a human investigated sample. The Canadian Hansard data set made available to us was de-accented, for example, 'é' and 'e' were both represented as 'e'. This resulted in a loss of information and the need to convert the Amnesty International press releases into a similar form. The accuracy loss produced in this process was small but non-zero. Overall, the average level of selection of the cross-language mate as the most similar document was approximately 80%.

The right bar in Figure 4 indicates the results when a random sample of 1,241 paired Amnesty International press releases was used to train the system and a random set of 2,482 Canadian Hansard documents made up of 1,241 conceptually identical documents in each language was used as the

test set. This process was repeated six times to achieve an overall average performance of approximately 63% accurate pairings. This shows that there is a significant fall in performance when documents from one corpus are used to train the system and it is tested on documents from a different corpus.

5 Conclusions and further work

Unsurprisingly, using a background set from a different corpus to the test set degrades both LSA and CL-LSA performance. LSA falls from being only slightly less similar to the average human than a typical person to being very much less similar. This indicates that it is important to train LSA on a set of documents that contains similar concepts to those in the document set that it is to be used upon. Removing stop words, performing log-entropy weighting of the term-document matrix and retaining over 500 factors were found to be the optimal parameters for LSA adding further support to the optimality of these parameters. The fall in the performance of CL-LSA on a retrieval task is in the order of 20% but still results in the situation where around 60% of documents are matched with their exact translation being considered the most similar document to them. This is still an acceptable level of accuracy under a broad range of conditions.

The difference between degradation in performance matching human similarity judgements and in matching documents to their exact translation is an interesting one. It indicates that CL-LSA is better at matching human performance on documents that are similar to each other than on those that are dissimilar. Again, this is a good feature as people typically want to retrieve documents that are most relevant to their search rather than those that are least relevant.

The mono-lingual application of LSA showed that the Matlab code used as a basis from which CL-LSA was implemented did not perform as well as the Perl, shell script and C code calling SVDPACK that had been used in

the past [8, 10]. While implementation of CL-LSA was easier from Matlab it may be that it could perform better when implemented using SVDPACK. Certainly an increase in speed may obviate the need to use volume reduction methods such as the "*¼doc*" scheme which reduce accuracy.

Comparison to human similarity judgements was used as the basis for comparison for LSA but no similar data set was available in a cross-lingual setting. Gathering such a dataset constitutes important work if we are to understand how well CL-LSA can model the similarity judgements of people and therefore how closely a set of documents returned from a search will match the set that would have been returned had it been performed by a person with enormous time and patience.

The only high quality translations available in many small languages are religious texts, particularly the Bible. However, Biblical language differs from Canadian Hansard much more than does that of Amnesty International press releases. Re-examining this process with the Bible used as the training corpus could provide a better indication of how CL-LSA would perform in reality.

The principals of CL-LSA should apply equally well to consonant based scripts such as Modern Standard Arabic or Hebrew, syllabics such as Japanese and ideographic scripts such as Chinese but the actual implementation will vary. Word separation in Chinese is a particularly hard problem that needs to be dealt with in CL-LSA implementation. Therefore implementation in non-Latin scripts is worthwhile and challenging future work.

It is possible to apply CL-LSA to situations where documents are available in more than two languages. For example, many Amnesty International press releases are now available in English, French, Spanish, Arabic and Russian and many United Nations documents are translated into their official languages of English, French, Russian, Chinese, Spanish and Arabic. The use of multiple languages would degrade performance and it would be interesting to measure this degradation.

Acknowledgements The Canparell data set was graciously provided by Prof. Tom Landauer, Praful Mangalath and the University of Colorado and the ABC Newsmail data set by Prof. Michael Lee and Dr. Matthew Welsh. Dr. Marcus Butivicius provided invaluable advice in the formulation of the article and fought to fund this work and employ the lead author. Lael Ferguson's help with Perl scripting made this work considerably easier. Dr. Garry Newsam made some very useful comments about SVD during the construction of this article that helped the authors understand the implications of the precipitous drop in singular values. We thank the vacation studentship program at DSTO for funding William Cox's time on this work and DSTO's ARM07/165 task for funding its completion.

References

[1] K. Börner. Extracting and visualizing semantic structures in retrieval results for browsing. In Peter J. Nürnberg, David L. Hicks and Richard Furuta, editors, *Proceedings of the fifth ACM conference on Digital libraries*, pages 234–235. ACM 2000. doi:http://doi.acm.org/10.1145/336597.336672 C1055

[2] Deerwester, S. C., Dumais, S. T., Landauer, T. K., Fernas, G. W. and Harshman, R. A., Indexing by Latent Semantic Analysis, *Journal of the American Society of Information Science*, **41**, 1990, 391–407. doi:10.1002/(SICI)1097-4571(199009)41:6¡391::AID-ASI1¿3.0.CO;2-9 C1063, C1069

[3] S. T. Dumais, T. K. Landauer and M. L. Littman. Automatic cross-linguistic information retrieval using Latent Semantic Indexing. In *SIGIR'96 - Workshop on Cross-Linguistic Information Retrieval*, pages 16–23. ACM, 1996. C1055, C1056, C1057, C1059, C1069

[4] T. K. Landauer and M. L. Littman. Fully automatic cross-language document retrieval using latent semantic indexing. In Gregory Grefenstette, editor, *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, pages 31–38. UW Centre for the New OED and Text Research, Waterloo Ontario, 1990. C1057, C1058, C1069

[5] Landauer, T. K., Littman, M. L. and Stornetta, W. S., A statistical method for cross-language information retrieval. Unpublished manuscript, 1992. C1055, C1059, C1069

[6] Landauer, T. K., Foltz, P. W. and Laham, D., Introduction to Latent Semantic Analysis, *Discourse Processes*, textbf25, 1998, 259–284. C1055, C1069

[7] Lloyd, R. and Shakiban, C., Improvements in Latent Semantic Analysis, *American Journal of Undergraduate Research*, **3**, 2004, 29–34. http://www.ajur.uni.edu/v3n2 C1055

[8] B. Pincombe. Comparison of Human and Latent Semantic Analysis (LSA) Judgements of Pairwise Document Similarities for a News Corpus. *Research Report* DSTO-RR-0278. DSTO, 2004. http://dspace.dsto.defence.gov.au/dspace/bitstream/1947/3334/1/DSTO-RR-0278%20PR.pdf C1060, C1062, C1065, C1066, C1069, C1071

[9] P. G. Young. *Cross-language information retrieval using latent semantic indexing*. Technical Report UT-CS-94-259. University of Tennessee, 1994. C1057

[10] M. D. Lee, B. M. Pincombe and M. B. Welsh. An empirical evaluation of models of text document similarity. In Bruno G. Bara, Lawrence Barsalou and Monica Bucciarelli, editors, *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 1254–1259. Lawrence Erlbaum Associates, Mahwah, NJ, 2005.

http://hdl.handle.net/2440/28910 C1056, C1057, C1065, C1066, C1069, C1071

Author addresses

1. **W. Cox**, Intelligence, Surveillance and Reconnaissance Division, Defence Science and Technology Organisation, PO Box 1500, Edinburgh, South Australia 5111, Australia.
mailto:william.cox@student.curtin.edu.au

2. **B. Pincombe**, Land Operations Division, Defence Science and Technology Organisation, PO Box 1500, Edinburgh, South Australia 5111, Australia.