SANDIA REPORT
SAND2007-2706
Unlimited Release
Printed May 2007

Cross-Language Information Retrieval Using PARAFAC2

Peter A Chew, Brett W Bader, Tamara G Kolda, Ahmed Abdelali

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

Cross-language information retrieval using PARAFAC2

Peter A Chew, Brett W Bader, Tamara G Kolda
Sandia National Laboratories
Albuquerque, NM 87185, and Livermore, CA 94551, USA
+1 (505) 845-0011
{pchew, bwbader, tgkolda}@sandia.gov

Ahmed Abdelali
New Mexico State University
Las Cruces, NM 88003, USA
+1 (505) 646-5711
ahmed@crl.nmsu.edu

ABSTRACT

A standard approach to cross-language information retrieval (CLIR) uses Latent Semantic Analysis (LSA) in conjunction with a multilingual parallel aligned corpus. This approach has been shown to be successful in identifying similar documents across languages - or more precisely, retrieving the most similar document in one language to a query in another language. However, the approach has severe drawbacks when applied to a related task, that of clustering documents 'language-independently', so that documents about similar topics end up closest to one another in the semantic space regardless of their language. The problem is that documents are generally more similar to other documents in the same language than they are to documents in a different language, but on the same topic. As a result, when using multilingual LSA, documents will in practice cluster by language, not by topic.

We propose a novel application of PARAFAC2 (which is a variant of PARAFAC, a multi-way generalization of the singular value decomposition [SVD]) to overcome this problem. Instead of forming a single multilingual term-by-document matrix which, under LSA, is subjected to SVD, we form an irregular three-way array, each slice of which is a separate term-by-document matrix for a single language in the parallel corpus. The goal is to compute an SVD for each language such that V (the matrix of right singular vectors) is the same across all languages. Effectively, PARAFAC2 imposes the constraint, not present in standard LSA, that the 'concepts' in all documents in the parallel corpus are the same regardless of language. Intuitively, this constraint makes sense, since the whole purpose of using a parallel corpus is that exactly the same concepts are expressed in the translations.

We tested this approach by comparing the performance of PARAFAC2 with standard LSA in solving a particular CLIR problem. From our results, we conclude that PARAFAC2 offers a very promising alternative to LSA not only for multilingual document clustering, but also for solving other problems in cross-language information retrieval.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Conference'04, Month 1–2, 2004, City, State, Country.
Copyright 2004 ACM 1-58113-000-0/00/0004…$5.00.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *clustering, retrieval models.*

H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *performance evaluation (efficiency and effectiveness).*

General Terms

Algorithms, Measurement, Design, Experimentation, Languages, Theory, Verification.

Keywords

Latent Semantic Analysis (LSA), information retrieval, multilingual, clustering, PARAFAC2.

1. INTRODUCTION

As the World Wide Web (WWW) has developed, content has become readily available in a multitude of languages, and interest has grown in the problem of cross-language information retrieval (CLIR) (see for example [21]). Based on our own fairly informal survey (using Google and limiting results of a variety of queries by language), we believe that Figure 1 is a reasonable estimate of the distribution of internet content by language.

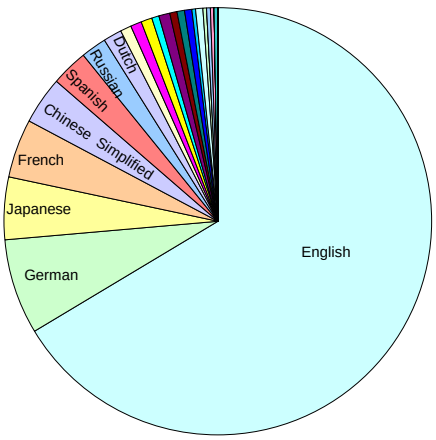

Figure 1. Estimated WWW content, distribution by language

Moves in various parts of the world towards political integration are another significant driver for the interest in CLIR. Nowhere is this more evident than in the European Union (EU), where official

documents are created, and must be managed, in an ever-increasing number of languages. Indeed, the EU has funded a significant amount of research in recent years into CLIR; the Cross-Language Evaluation Forum (CLEF) [21] is one example.

Our own interest in CLIR is as a means to cluster documents from the WWW. Clearly, these documents could be in any language, but we would like to cluster the documents by topic, factoring language out, so that documents on the same topic appear close to one another irrespective of their language.

In section 2, we outline a standard approach to CLIR, and in section 3, we describe our implementation of that approach. As described in section 4, we found that we were able to confirm that this approach worked well for certain CLIR problems, but that it had severe drawbacks when we attempted to use it for cross-language clustering. The reasons for this are discussed, and we propose a novel alternative approach using PARAFAC2 instead of standard SVD in section 5. We compare how PARAFAC2 measures up to standard LSA in practice in section 6, and conclude on our results in section 7.

2. A STANDARD APPROACH TO CLIR

A standard approach to cross-language information retrieval uses Latent Semantic Analysis (LSA) [11] in conjunction with a multilingual parallel aligned training corpus. This application of LSA to multilingual data is described in [5] and used in [23]. A term-by-document matrix of weighted frequencies is formed from the corpus; each 'document' consists of the concatenation of all the languages, so terms from all languages will appear in any given document. A variety of weighting schemes can be used, but the log-entropy weighting scheme is generally believed to be one of the most effective for this purpose [10]. In this scheme, the weighted frequency $x_{t,n}$ of a particular term t in a particular document n is given by:

$$x_{t,n} = \log_2 (F_t + 1) \cdot (1 + H_t / \log_2 (N))$$

where F_t is the raw frequency of t in n, H_t is a measure of the entropy of the term across all documents, and N is the number of documents in the corpus. (Accordingly, $\log_2 (N)$ is the maximum entropy that any term can have in the corpus; $(1 + H_t / \log_2 (N))$ is 1 for the most distinctive terms in the corpus, 0 for those which are least distinctive.)

In the standard approach, the term-by-document matrix of weighted frequencies X is subjected to SVD: $X = USV^T$. The output is a term-by-concept matrix (U, or the matrix of left singular vectors), a set of singular values (S, a diagonal matrix), and a document-by-concept matrix (V, or the matrix of right singular vectors). The number of columns computed for U and V is referred to as the number of LSA dimensions. Vectors for new documents (those not in the original parallel corpus) are computed by multiplying the vectors of weighted frequencies of terms in the new documents by US^{-1}. The cosine between any two such vectors is a measure of the similarity between those two documents.

There are a number of well-understood practical advantages to using an approach like LSA for CLIR. Essentially, the parallel corpus used for training acts like a 'Rosetta Stone'; it is the key which unlocks the door to comparing documents across language boundaries, while the underlying algorithms remain constant regardless of which languages are being compared. This becomes particularly advantageous when language-specific expertise is in short supply. An alternative approach to CLIR which is commonly employed, for example, is to translate documents: before computing a similarity, the source document is translated into the language of the target document. However, even if a machine translation (MT) system is used to automate this procedure, it is usually the case that a separate MT system must be put in place for each language pair, and that some familiarity with each language in the pair is required to build each such system. For any significant number of languages, the cost of building the required 'system of systems' is likely to be prohibitive, even if the expertise and resources required to do so are available. Another alternative approach (exemplified in [18]) is to use bilingual dictionaries, but these may not be available in all languages. In light of this, it is easy to see the attractiveness of a generic approach like LSA which relies only on the ability to tokenize text at the boundaries between words, or more generally semantic units – a procedure which can be generalized to virtually all languages, even logographic languages like Chinese.

3. IMPLEMENTATION OF THE STANDARD APPROACH

In implementing multilingual LSA, perhaps the major decision to be made is which parallel aligned corpus to use in training. For the work described here, we used the Bible. Although it is hard to come by reliable statistics which allow direct comparison, the Bible is generally believed to be the world's most widely translated book ([8], [9], [22]) with at least partial translations into at least 2,426 languages and full translations into at least 429 languages [6]. A single website [7] has at least 80 parallel translations in over 50 languages (Table 1 lists most of these); almost all of the translations available for download are public-domain, and all are in a tab-delimited format which can easily be aligned by verse (see Figure 2 for an example).

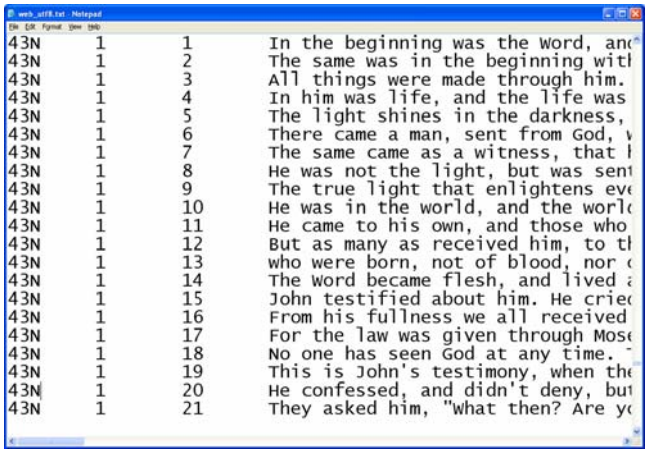

Figure 2. Sample data from publicly-available parallel corpus

Table 1. Languages potentially available for multilingual LSA

Language	Number of translations available
Afrikaans	1
Albanian	1
Arabic	1
Aramaic	1
Armenian (Eastern)	1
Armenian (Western)	1
Basque	1
Breton	1
Chamorro	1
Chinese (Traditional)	2
Chinese (Simplified)	2
Croatian	1
Czech	4
Danish	1
Dutch	1
English	8
Esperanto	1
Estonian	1
Finnish	2
French	2
German	5
Greek (Modern)	1
Greek (New Testament)	6
Hebrew (Modern)	1
Hebrew (Old Testament)	3
Hungarian	1
Indonesian	1
Italian	2
Japanese	1
Korean	1
Latin	1
Latvian	1
Lithuanian	1
Manx Gaelic	1
Maori	1
Norwegian	1
Polish	1
Portuguese	1
Romani	1
Romanian	1
Russian	1
Scots Gaelic	1
Spanish	3
Swahili	1
Swedish	1
Tagalog	1
Thai	1
Turkish	1
Vietnamese	1
Wolof	1
Xhosa	1
TOTAL	**80**

Conveniently for our purposes, all of the languages represented most frequently in the WWW (see Figure 1) are also represented in [7]. The list of represented languages is less biased towards European languages (or at least languages of a particular language group) than is commonly the case with purpose-built parallel corpora, a reflection of the reasons that the translations of the Bible exist in the first place.[1] In addition, there is evidence that even when sections of the parallel corpus are defective (for example, when only a portion of the Bible exists in a particular language), the defective sections can still be used without overall detriment [8]. We estimate, therefore, that using the Bible (in the dozens of translations that we have already downloaded) as a parallel corpus for training LSA, we would achieve about 99.75% coverage of internet content, a coverage which would have been hard to match using parallel text from any other single source.

A question which is commonly raised is how representative the vocabulary of the Bible is of modern vocabulary, and therefore how suitable it is as training data. One answer to this is that it depends on which translations are used; many languages have multiple translations of the Bible (among our downloads, for example, there are 8 English translations ranging from the King James Version, dating from 1604, to the World English Bible, dating from 2006). Clearly, the more modern the translation, the better will be the coverage of the modern language. According to [22], the Bible's coverage may be somewhere between 75%-85%, the vocabulary which is not covered consisting mostly of technical terms and proper names. Our own informal tests confirmed that this estimate is probably not too far off; based on a sample of 602,995 web pages we collected, and after removing items which were treated as words by our tokenizer but cannot reasonably be considered words (such as '^MF', '_G'), we believe that our coverage of *vocabulary* (as opposed to *languages*) from the WWW would be around 70%. In any case, there is no reason to suppose that coverage has to approach 100% to allow for effective CLIR: in fact, we shall present evidence in this paper that vocabulary coverage of even less than 60% is sufficient to allow a high level of precision in solving certain CLIR problems. And although we have used the Bible as the training data, there is no reason that the approach could not be extended to the Bible *plus* additional parallel corpora.

Since the Bible is alignable by verse, and there are more than 30,000 verses in the Bible, each averaging about a sentence or two in length, an extremely fine-grained term-by-document matrix can be created. Generally, we have found that the finer the granularity, the better CLIR results we obtain. With 77 parallel versions and using our alignment scheme[2], our term-by-document

[1] By contrast, parallel corpora developed with government funding, for example, are understandably more restricted in scope. The corpora developed with EU funding, for example, naturally consist mostly of material in European languages.

[2] 77 is the greatest number of parallel versions we have used so far in multilingual LSA. The alignment of the raw data in [7] is not always perfect owing to minor differences in versification between translations. We addressed this by spending some time cleaning the raw data to improve the alignment. As a result, the number of verses in our alignment scheme (31,226) is greater than the number of verses in most of the raw downloads (31,102).

matrix was 1,454,289 by 31,226. As is typical in natural language processing, this matrix is extremely sparse; the number of nonzeros in this case was 21,759,766, representing a density of around 0.048%. We stored the parallel text in a relational SQL Server database to allow for easy aggregation of the statistics required to form different term-by-document matrices for different language combinations and use by different CLIR algorithms. To compute the SVD, we used either SVDPACK [4] or a library called Anasazi [3], which is part of the Trilinos framework [14]. In each case, we computed a truncated SVD corresponding to the 300 highest singular values. We found, however, that SVDPACK was unable to cope with the size of term-by-document matrices necessary to process more than around two dozen languages in parallel, and thus we resorted in these cases to using Trilinos (which is designed to run on a Linux cluster and is consequently considerably more scalable). The results of SVD were then imported back into SQL Server and we used SQL scripts to compute the vectors for new documents or queries, for example those in the test set.

4. VALIDATION OF LSA

4.1 Test data and method

The test data we used were the 114 suras (chapters) of the Quran, which has also been translated into a number of languages. Clearly, test data of this sort are a prerequisite in order to be able to measure effectiveness in multilingual clustering. For most of the work described in this paper, we limited the selection of languages to Arabic, English, French, Russian and Spanish (the respective abbreviations AR, EN, FR, RU and ES are used hereafter), in both the training and the test data. With this data, the initial term-by-document matrix was 160,396 by 31,226 with 2,684,938 nonzeros. With the five languages, the test data amounted to 570 documents: a relatively small set, but large enough to achieve statistically significant results for our purposes, as will be shown. Note also that although the test documents all come from a single topic domain, it is reasonable to assume that the *comparative* results we will report in this paper are valid in general, because in all tests we describe, we are using the same test set.

Perhaps surprisingly, the Bible's coverage of the Quran's vocabulary appears to be lower than the Bible's coverage of general WWW vocabulary. Of 58,015 distinct terms in the Quran, only 33,423 (or about 58%) appear in the Bible.

We tokenized each of the 570 test documents, applying the weighting scheme described above to obtain a vector of weighted frequencies of each term in the document, then multiplying that vector by $U \times S^{-1}$, also as described above. The result was a set of projected document vectors in the 300-dimensional LSA space.

4.2 Evaluation measures

We used four separate measures to evaluate the effectiveness of CLIR given this data. These measures are listed in Table 2.

Table 2. CLIR measures

#	Measure
1	Precision at 1 document (for a given source and target language)
2	Precision at 0 (for a given source and target language)
3	Multilingual precision at 5 documents (for 5 languages)
4	Multilingual precision at 0 (for 5 languages)

There is sometimes confusion about the different measures of precision, so for the avoidance of all doubt, we shall spell out how exactly these measures are calculated. The first of these, precision at 1 document, is the proportion of cases, on average, where the translation was retrieved first. For example, if French sura number 5 was the most similar sura among all the French suras to English sura number 5, then precision at 1 document in this case would be 1, and 0 otherwise. This is a strict measure, since no more credit is given if the translation is ranked second than if it is ranked bottom. The second measure, precision at 0, is less strict. This represents the maximum precision at any level of recall. Since we are dealing with translations, only one document is considered relevant, and precision at 0 is therefore the inverse of the ranking of the translation. These first two measures relate to the effectiveness of our CLIR technique in finding similar documents *given the language of the query and the language of the results*, and for convenience we will refer to these two measures collectively as 'language-specific' precision metrics.

Measures 3 and 4, on the other hand, relate to the effectiveness of our technique in finding similar documents *regardless of source or target language*. These measures give an indication of how well multilingual clustering is likely to work. Since we have 5 languages, the best result we could achieve for clustering would be to have all five translations ranked in the top 5 in similarity to the query. 'Multilingual precision at 5 documents', therefore, represents the proportion of the top 5 retrieved results which are translations of the query, and 'multilingual precision at 0' (again, the less strict measure) represents the maximum precision at any level of recall after the fifth document. We refer to these two measures as 'multilingual' precision metrics.

4.3 Results with LSA

Using the standard approach and measures 1 and 2 as the evaluation metric, we obtained our best results using LSA with the given languages with 280 dimensions. These results are shown in Table 3 and Table 4.

Table 3. Precision at 1 document with standard LSA

	AR	EN	ES	FR	RU
AR	1.000	0.500	0.491	0.570	0.474
EN	0.684	1.000	0.912	0.974	0.833
ES	0.500	0.860	1.000	0.930	0.605
FR	0.605	0.930	0.947	1.000	0.789
RU	0.474	0.825	0.798	0.789	1.000

Table 4. Precision at 0 with standard LSA

	AR	EN	ES	FR	RU
AR	1.000	0.656	0.653	0.695	0.645
EN	0.765	1.000	0.935	0.983	0.899
ES	0.630	0.897	1.000	0.953	0.731
FR	0.711	0.961	0.964	1.000	0.869
RU	0.608	0.877	0.866	0.854	1.000

On average, precision at 1 document here is 0.780, and precision at 0 is 0.846. (As more parallel translations are added, both these precisions rise further, to around 0.81 and 0.87 with 52 languages and 77 parallel translations [8].) These averages include the diagonal values of 1.000. These reflect very favorably on the ability of the standard approach to identify translations, providing the search space is limited in each case to a single language: here, almost 80% of the time, the translation is retrieved first. The results also compare favorably with published results which use different methodologies for CLIR (using a different data set, McNamee and Mayfield report mean average precision of no more than 0.45 for English-to-Spanish CLIR using 5-grams [19]). Recall that these results were achieved despite the Bible's coverage of the Quran's vocabulary being less than 60%; proof, it would seem, that even with only partial coverage of the target vocabulary, CLIR can be very effective.

Under measures 3 and 4, however, a different picture emerges. The relevant results are presented in Table 5 (not broken down by language pair, because the different languages are now mixed together in the test set).

Table 5. 'Clustering' precision with standard LSA

Measure	Results
Multilingual precision at 5 documents	0.259
Multilingual precision at 0	0.265

It is worth noting that under standard LSA, while language-specific precision tends to increase as more LSA dimensions are used (at least up to 300 dimensions, which is as far as we have tested), the opposite seems to be true for multilingual precision, at least to a certain point. Above 5 dimensions, it appears that multilingual precision generally decreases (see Figure 3).

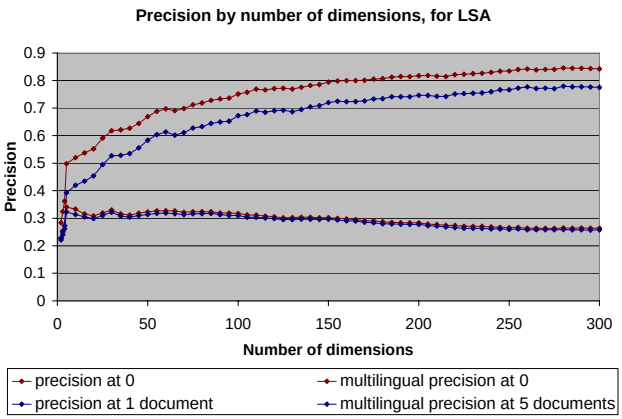

Figure 3. Effect of number of dimensions on LSA

Following the findings in [8], we attempted to boost either language-specific or multilingual precision by increasing the number of parallel translations used in training LSA. Our results did confirm the premise in [8] that more linguistic parallelism is beneficial to LSA (both for language-specific and multilingual precision). However, even with 77 parallel translations, multilingual precision rises no higher than 0.300 (at 5 documents) and 0.307 (at 0); and from Figure 3 it will be seen that in the best case, we were unable to achieve multilingual precision at 0 of above around 0.35 using LSA. Considering that these measures can never be lower than 0.2 with 5 languages (since each document is always most similar to itself, and therefore ranks top in the results), these results are all the more unimpressive: on average, hardly any of the second-to-fifth ranked results are translations of the query. How can this be, when the first two measures produced much more encouraging results?

4.4 Limitations of LSA

In part, this can be answered by considering what happens when we attempt to use the LSA document vectors to 'map' the documents in a graphical representation, such that similar documents are located close to one another. When we attempted this, we found that the documents essentially cluster by language, not by topic. To understand how this can happen, consider the hypothetical example of some ranked results shown in Table 6.

Table 6. Example illustrating low multilingual precision

Ranking	Language of retrieved document	Relevant?
1	English	Yes
2	English	No
3	English	No
4	English	No
5	English	No
6	French	Yes
7	French	No
8	Spanish	Yes
9	Russian	Yes
10	Arabic	Yes

In this example, while the first document retrieved in each language was the relevant one, many non-relevant English documents were returned before the relevant documents in the other languages. In this example, measures 1 and 2 would each have been 1, but measure 3 would have been only 0.2. This occurs, we believe, because each language has its own distinctive statistical 'signature', as is reflected in the differing counts for 'types' (unique terms) versus 'tokens' (instantiations of those terms in the text) in the parallel Bible text we used in training. These counts are shown in Table 7.

Table 7. Types and tokens in the Bible by language

	Types	Tokens
Arabic	55,300	440,435
English	12,335	789,744
French	20,428	812,947
Russian	47,226	560,524
Spanish	28,456	704,004
TOTAL	163,745	3,307,654

Assuming that the translations of the Bible in our parallel corpus are accurate and complete, this table would appear to suggest that Arabic takes just over half the number of terms to express the same amount of information as English, that English and French take similar numbers of terms, and so on. Intuitively, this seems right given that Arabic and Russian rely much more than than English, French and Spanish on the use of morphology (endings, and so on) to add to or modify the meanings of words. The same phenomenon can also be illustrated well on a small scale by considering the first 'document' in the parallel corpus (the first verse), shown in Table 8. It is likely to be no coincidence that the best cross-language prediction results we achieved were for pairs of languages with similar statistics (for example, English and French), and the worst results were for those with dissimilar statistics (such as Arabic and English) (see Table 3 and Table 4).

Table 8. Illustration of statistical differences between languages

	Text	word count	% of total
AR	في البدء خلق الله السموات والارض.	6	14
EN	In the beginning God created the heavens and the earth.	10	24
FR	Au commencement Dieu créa les cieux et la terre.	9	21
RU	В начале сотворил Бог небо и землю.	7	17
ES	En el principio crió Dios los cielos y la tierra.	10	24
TOTAL		**42**	**100**

The statistical differences can, in fact, be shown to have a detrimental effect on LSA – not just empirically, but theoretically as well. Under the standard log-entropy weighting scheme, we can verify whether, according to this scheme, the contribution of each of the 5 languages in our multilingual aligned parallel corpus is equal – which it should be, if the translations are complete and accurate. The computed entropy is a measure of information content: because the same information is being conveyed in the translations of any given document in the training corpus, the total entropy per language (the sum of term entropies of terms in that language) should be constant for any given document.

Upon examination, we found that with the standard log-entropy weighting the computed information content varies quite widely by language, which is perhaps unsurprising. If it takes 560,524 Russian words to express what English says in 789,744 words, then on average Russian words must contain more information (or meaning) than English words (again, a notion which is consistent with what we know about the way words are formed in Russian and English). However, since entropy (or information content) in the log-entropy scheme is simply the entropy of a particular term across all documents, and since the scheme takes no account of the specific properties of different languages, there is no guarantee that the contributions of different languages in the parallel corpus will be equal as they should be. In fact, in our parallel corpus, where under LSA all languages are 'mixed together' in the bag-of-words approach, languages which have more terms overall (such as English and French) generally account for a higher percentage of the 'information' in each document. This points to a flaw in standard multilingual LSA, or at least in the log-entropy weighting scheme as applied within that approach.

One other point to note is the difference between the total of 163,745 shown in Table 7 above, and the figure of 160,396 mentioned in section 4.1. The difference of 3,349 represents those terms that occur in more than one language, such as 'de' ('of' in French and Spanish), English 'coin' versus French 'coin' ('corner'). The relatively small number of such terms is unlikely to affect the cross-language precision results significantly, but it is worth pointing out that standard LSA has no way to distinguish between homographs from different languages, and in some cases this could be problematic, especially when the homographs have very different meanings in the different languages.[3]

Given all this, the statistical explanation seems to be a reasonable one for why, when we attempted to map the documents graphically such that similar documents were close to one another, the documents clustered by language rather than by topic. Precisely the same issue has been identified elsewhere in the literature: Mathieu et al [18] report that 'even if the cross-lingual similarity measure is designed to behave the same when comparing documents written in the same language and documents written in different ones, our evaluation shows that it still tends to gather in a cluster documents of same language prior to different language ones'.

5. AN ALTERNATIVE APPROACH

As discussed in the previous section, it is a drawback of the standard approach to LSA that there is no delineation between different languages in the training data. All languages are concatenated together in training, so that each 'document' is multilingual. Within the LSA framework, however, this is unavoidable, since without the concatenation, LSA is unable to make the associations between words in different languages when they co-occur.

[3] The difference between the total of 3,307,654 and the 2,684,938 nonzeros mentioned in section 4.1 can also be explained: the figure in Table 6 is higher because some terms occur more than once in the same verse. For example, if 'the' occurs three times in a particular verse, this would account for three tokens, but only one nonzero entry in the term-by-document matrix.

To overcome this problem, therefore, we propose a novel application of PARAFAC2 [13] as an alternative to LSA. PARAFAC2 is a variant of PARAFAC [12], a multi-way generalization of the SVD. The PARAFAC model is based on a 'parallel proportional profile principle' that applies the same factors across a parallel set of matrices to minimize a least-squares objective. Let the M × N matrix X_k, k = 1, …, K, denote the *k*th slice of a three-way data array X, and let R be the number of dimensions of the LSA conceptual space. Then the standard PARAFAC model is

$$X_k = U S_k V^T \quad (1)$$

where U is an M × R factor matrix for the terms, S_k is an R × R diagonal matrix of weights for the kth slice of X, and V is an N × R factor matrix for the documents. In this form, it is easy to see PARAFAC's similarity to the SVD. Here, though, we find factor matrices U and V that are the same for every matrix X_k. However, the factors U and V are not orthogonal as they are for the SVD.

In our application, we can let X_k be the term-by-document matrix for the *k*th language in the parallel corpus. It has M_k terms and N documents; however, since the number of rows in each slice differs, the PARAFAC model is not appropriate. PARAFAC2 is a related model that is appropriate because it relaxes the constraint that the U matrix is the same across all slices. Thus, we form an irregular three-way array, each slice of which is a separate term-by-document matrix for a single language in the parallel corpus. The number of documents in each slice will be the same, since the corpus is parallel, but the number of terms will vary by language. The K=5 slices of X for our application are shown in Figure 4.

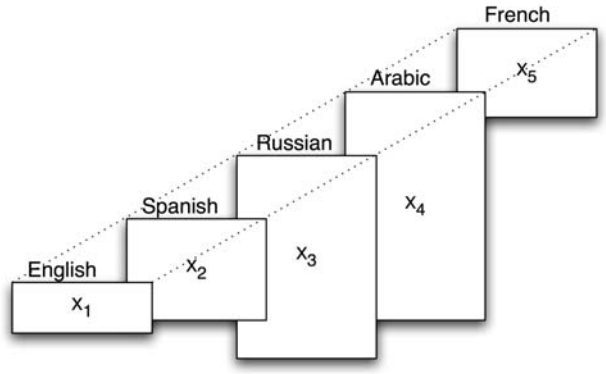

Figure 4. Term-by-document matrices by language as a three-way array (X)

The PARAFAC2 has the following form:

$$X_k = U_k H S_k V^T \quad (2)$$

Here, there is an orthonormal M_k × R factor matrix U for each slice of X, and an H matrix of size R × R. Because this model lacks certain uniqueness properties associated with the standard PARAFAC model, an invariance constraint is needed on the left factor matrices (i.e., the product U_kH). To gain uniqueness, Harshman [12], [13] imposed the constraint that the cross product $(U_kH)^T(U_kH)$ is constant over k, which in this formulation is accomplished with the constraint that H is nonsingular. The PARAFAC2 model is shown in Figure 5.

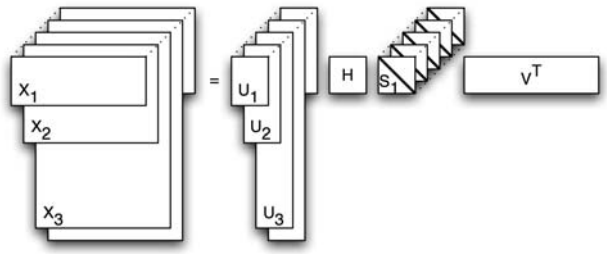

Figure 5. The PARAFAC2 model

Conceptually, the goal is to compute something like an SVD for each language such that V (analogous to a matrix of right singular 'document' vectors, though not orthonormal) is the same across all languages, although for each language k there will be a separate U_k (analogous to a matrix of left singular 'term' vectors for language k) and S_k (analogous to singular values). A benefit of PARAFAC2 is that it has a separate mapping for each language into the LSA conceptual space; in particular, each mapping is orthogonal for each language rather than the one large orthogonal mapping for all languages at once. In other words, PARAFAC2 imposes the constraint, not present in standard LSA, that the 'concepts' (i.e., columns of U_k) of any given language in the parallel corpus taken on its own map to those of any other language. Intuitively, this constraint makes sense, since the whole purpose of using a parallel corpus is that translations are supposed to render the same concepts in different languages.

To compute the PARAFAC2 model of X, we implemented a variant of the algorithm outlined in [15] that is adapted to handle very large and sparse data. The complete procedure is summarized as follows.

Step 0. Initialize V as the R principal eigenvectors of $\sum_k X_k^T X_k$ and initialize H and $S_1,...,S_k$ as R × R identity matrices.

Step 1a. Compute the SVD of $Z_k = HS_kV^TX_k^T = P_k\Sigma_kQ_k$ by first computing the R principal eigenvectors of $Z_kZ_k^T$ to obtain P_k and normalizing the columns of $Z_k^TP_k$ to obtain Q_k, and then update U_k as $Q_kP_k^T$, k = 1,...,K.

Step 1b. Update H, V, and $S_1,...,S_k$ by one iteration of an alternating least squares algorithm for standard PARAFAC, equation (1), applied to the R x N x K three-way array with frontal slices $U_k^TX_k$, k = 1,...,K. (See [1]0 for an efficient implementation with large-scale data.)

Step 2. Repeat step 1 until a maximum number of iterations has been reached or the norm of the residual, $\sum_k \|X_k - U_k H S_k V^T\|$, ceases to change appreciably.

This algorithm was written in MATLAB using the Tensor Toolbox [1], [2], and the PARAFAC2 model was computed on a dual 3GHz Pentium Xeon desktop computer with 2GB of RAM.

Once the PARAFAC2 model has been computed for all languages according to these constraints, the manner in which new documents are projected into the semantic space is similar to that used in LSA. A vector of weighted term frequencies (the term-by-document vector) is formed as described in 4.1 above. The difference is that this vector is multiplied by the $U_kS_k^{-1}$ specific to the language of the document, rather than the general US^{-1} for all

languages which is the artifact of LSA. This relies, of course, on knowing the language of the new document, but there are a variety of machine learning methods for reliably determining the language of an unseen document; one such method (which achieves an accuracy of over 99%) is mentioned in [20], and we have achieved similar results by training a neural network on the LSA vectors. Thus, it can be seen that the additional step necessitated by PARAFAC2 could easily be automated and is not a significant obstacle to wider deployment.

The main disadvantage of PARAFAC2 compared to LSA is that more computation is required to obtain the decomposition. In fact, since there is currently no parallel implementation of PARAFAC2, we can compute at most 240 dimensions using PARAFAC2. However, as with LSA, the PARAFAC2 decomposition need only be computed once, and the results are then available for use multiple times, so the one-time cost of using PARAFAC2 is essentially one which can be highly leveraged.

This disadvantage in performance is also offset by an advantage which applies at run-time: since the language-specific U_k matrices are considerably smaller than the general U matrix, the process of matrix multiplication can be considerably faster than it is under LSA. There is another linguistic/theoretical advantage to PARAFAC2, and this has to do with the 'homographs' issue identified in section 4.4 above. Since, under PARAFAC2, we are now delineating between the input of different languages in training, English 'coin' is differentiated from French 'coin' – which, one would assume, can only be advantageous in CLIR since the homographs in this particular pair are, as far as we know, unrelated in meaning.

6. RESULTS USING PARAFAC2

With the same training and test data as described in section 4 above, and using PARAFAC2, we obtained the results shown in Table 9, Table 10, and Table 11. Since we were limited to 240 dimensions, for a fair comparison we also recalculated precision under LSA using only the top 240 dimensions. The relevant results are shown in Table 12, Table 13, and Table 14.

Table 9. Precision at 1 document with PARAFAC2

	AR	EN	ES	FR	RU
AR	1.000	0.667	0.693	0.746	0.693
EN	0.632	1.000	0.947	0.982	0.833
ES	0.605	0.947	1.000	0.974	0.886
FR	0.728	0.974	0.956	1.000	0.895
RU	0.728	0.921	0.895	0.939	1.000

Table 10. Precision at 0 with PARAFAC2

	AR	EN	ES	FR	RU
AR	1.000	0.785	0.793	0.827	0.793
EN	0.738	1.000	0.968	0.990	0.887
ES	0.705	0.967	1.000	0.981	0.918
FR	0.791	0.989	0.972	1.000	0.930
RU	0.807	0.947	0.935	0.958	1.000

Table 11. 'Clustering' precision with PARAFAC2

Measure	Results
Multilingual precision at 5 documents	0.402
Multilingual precision at 0	0.415

Table 12. Precision at 1 document - LSA, 240 dimensions

	AR	EN	ES	FR	RU
AR	1.000	0.447	0.456	0.579	0.561
EN	0.649	1.000	0.904	0.965	0.746
ES	0.465	0.798	1.000	0.921	0.596
FR	0.518	0.939	0.956	1.000	0.734
RU	0.439	0.754	0.798	0.763	1.000

Table 13. Precision at 0 – LSA, 240 dimensions

	AR	EN	ES	FR	RU
AR	1.000	0.600	0.607	0.704	0.678
EN	0.736	1.000	0.937	0.978	0.842
ES	0.591	0.859	1.000	0.946	0.727
FR	0.652	0.966	0.967	1.000	0.832
RU	0.585	0.845	0.856	0.845	1.000

Table 14. 'Clustering' precision – LSA, 240 dimensions

Measure	Results
Multilingual precision at 5 documents	0.261
Multilingual precision at 0	0.268

From these results it can be seen that PARAFAC2 outperforms standard LSA by a significant margin on the multilingual precision metrics – 0.402 compared to 0.261, or 0.415 compared to 0.268 depending on which measure is used. This is empirical confirmation that PARAFAC2 lives up to its promise, which is to ensure that the 'concepts' of the different languages are aligned with one another, and to factor out some of the statistical differences between languages that caused problems for LSA.

It is interesting to note that, based on this set of results, PARAFAC2 also appears to outperform LSA (by a narrower but still highly significant margin) in the language-specific metrics. The average precision at 1 document is 0.866 for PARAFAC2 compared with 0.760 for LSA, and for precision at 0 the averages are 0.907 and 0.830 respectively. Moreover, it will be seen by comparing Table 9 with Table 12, and Table 10 with Table 13, that the results using PARAFAC2 are superior almost across the board. The only exceptions are in precision at 1 document: English-to-Arabic was slightly lower for PARAFAC2, and French-to-Spanish was a tie. In all cases, precision at 0 is better under PARAFAC2. Since the average precisions represent the averages across 2,850 ($114 \times 5 \times 5$) query submissions, the differences between the results for PARAFAC2 and LSA are highly significant ($p \approx 5.22 \times 10^{-40}$ for overall average precision at 1 document, using a chi-squared test). We repeated the same comparisons at various different numbers of dimensions and found that PARAFAC2 consistently outperformed LSA, no matter how many dimensions the decomposition was computed in, and usually the difference was highly statistically significant. In fact,

even our best results using standard LSA[4] still could not compare with the PARAFAC2 results in Table 9 above.

For reference and comparison with Figure 3, the effect of the number of dimensions on precision under PARAFAC2 (to the extent we have run tests, and with lines to interpolate for numbers of dimensions not tested) is shown in Figure 6.

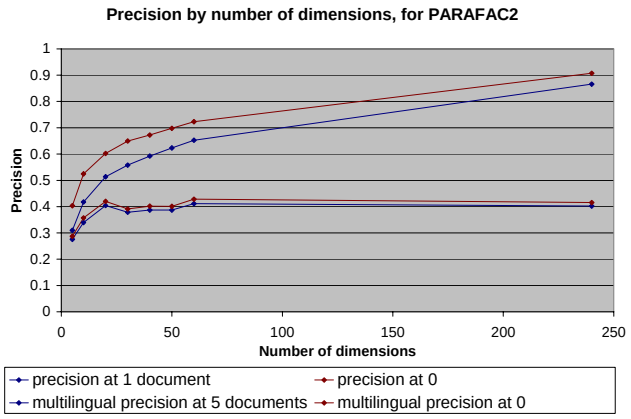

Figure 6. Effect of number of dimensions on PARAFAC2

It seems, therefore, that the effect of number of dimensions upon precision under PARAFAC2 follows a pattern similar to that for LSA.

7. CONCLUSION

In summary, the evidence appears to be highly compelling that PARAFAC2 is a superior alternative to standard LSA for multilingual information retrieval, at least for the two CLIR problems we want to solve. In line with our expectations, we found that this was particularly true for multilingual document clustering. However, since we had achieved respectable 'language-specific' results using LSA and thus already found it an effective tool for identification of translations, it was more unexpected for us to find that PARAFAC2 essentially beats LSA 'at its own game'. Even by the language-specific metrics which portray LSA in a good light, PARAFAC2 is a more effective tool than standard LSA.

In section 2, we outlined some of the qualitative features which make LSA attractive as a vehicle for CLIR: essentially, its extensibility to virtually all languages, particularly when used in conjunction with a widely-translated parallel corpus such as the Bible. It is important to note that all of these qualitative advantages apply just as much to PARAFAC2 as they do to LSA.

Although PARAFAC2 has a greater lead over standard LSA in the metrics which relate to multilingual clustering than it does in those that relate to language-specific CLIR, it has to be said that the initial baseline set by LSA was much lower (0.27 for multilingual precision at 0, compared with 0.83 for language-specific precision at 0). Further, even with the boost that PARAFAC2 provides for multilingual precision, the highest multilingual precision that we were able to attain (scarcely over 0.4) is not as high as we had hoped, and we are still doubtful that this level of precision will overcome the problem that we had hoped to solve, that of preventing documents from simply clustering by language in a graph-based analysis.

Nevertheless, PARAFAC2 represents a good step forward from LSA in addressing this problem. We intend to carry out further experiments to determine whether further adaptations can be made to PARAFAC2 to allow for multilingual document clustering to be carried out successfully. It remains to be seen what these adaptations might look like and to what extent we can streamline the method to maximize multilingual precision, but given the fact that our research with PARAFAC2 is still in a relatively initial stage, we are extremely optimistic that PARAFAC2 offers a promising way forward for truly language-independent clustering of documents by topic.

[4] With standard LSA, precision at 1 document averaged around 0.82 with 300 dimensions and 45 or more parallel translations used in training.

8. ACKNOWLEDGEMENTS

We are grateful to Steve Verzi, Stephen Helmreich, and Brad Mancke for the many constructive comments they have given us as we have worked on the material for this paper.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

9. REFERENCES

[1] Bader, B. W., and Kolda, T. G. Efficient MATLAB computations with sparse and factored tensors. Technical Report SAND2006-7592, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, Dec. 2006.

[2] Bader, B. W., and Kolda, T. G. MATLAB Tensor Toolbox, version 2.2. http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/, February 2007.

[3] Baker, C. G., Hetmaniuk, U. L., Lehoucq, R. B., and Thornquist, H. K. Anasazi: Block Eigensolver Package Web Site: http://software.sandia.gov/trilinos/packages/anasazi/, 2007.

[4] Berry, M. W., Do, T., O'Brien, G. Krishna, V., and Varadhan, S. *SVDPACKC (Version 1.0) User's Guide*. Knoxville, TN: University of Tennessee, 1996.

[5] Berry, M. W., Dumais, S. T., and O'Brien, G. W. Using Linear Algebra for Intelligent Information Retrieval. *SIAM: Review*, 37, 1994, 573-595.

[6] Bible Society. *A Statistical Summary of Languages with the Scriptures*. Accessed at http://www.biblesociety.org/latestnews/latest390-slr2006stats.html on February 27, 2007.

[7] Biola University. *The Unbound Bible*, 2005-2006. Accessed at http://www.unboundbible.com/ on February 27, 2007.

[8] Chew, P. A., and Abdelali, A. *Benefits of the 'Massively Parallel Rosetta Stone': Cross-Language Information Retrieval with over 30 Languages*, forthcoming.

[9] Chew, P. A., Verzi, S. J., Bauer, T. L., and McClain, J. T. Evaluation of the Bible as a Resource for Cross-Language

Information Retrieval. *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, 2006, 68–74.

[10] Dumais, S. T. Improving the Retrieval of Information from External Sources. *Behavior Research Methods, Instruments, and Computers* 23 (2), 1991, 229-236.

[11] Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S. and Harshman, R. Using Latent Semantic Analysis to Improve Access to Textual Information. In CHI'88: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1988, 281-285. ACM Press.

[12] Harshman, R. A. Foundations of the PARAFAC Procedure: Models and Conditions for an "Explanatory" Multi-Modal Factor Analysis. *UCLA Working Papers in Phonetics* 16, 1970, 1-84.

[13] Harshman, R. A. PARAFAC2: Mathematical and Technical Notes. *UCLA Working Papers in Phonetics* 22, 1972, 30-47.

[14] Heroux, M., Bartlett, R., Howle, V., Hoekstra, R., Hu, J., Kolda, T., Lehoucq, R., Long, K., Pawlowski, R., Phipps, E., Salinger, A., Thornquist, H., Tuminaro, R., Willenbring, J., Williams, A., and Stanley, K. An Overview of the Trilinos Project. *ACM Transactions on Mathematical Software* 31, No. 3, 2005, 397-423.

[15] Kiers, H. A. L., Ten Berge, J. M. F., and Bro, R. PARAFAC2 – Part 1. A Direct Fitting Algorithm for the PARAFAC2 Model. *Journal of Chemometrics* 13, 1999, 275-294.

[16] Kolda, T. G. and Bader, B. W. The TOPHITS model for web link analysis. In *Workshop on Link Analysis, Counterterrorism and Security*, 2006.

[17] Landauer, T. An Introduction to Latent Semantic Analysis. *Discourse Processes* 25, 1998, 259-284.

[18] Mathieu, B., Besançon, R. and Fluhr, C. Multilingual Document Clusters Discovery. *Recherche d'Information Assistée par Ordinateur (RIAO) Proceedings*, 2004, 1-10.

[19] McNamee, P. and Mayfield, J. Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval* 7, 2004, 73-97.

[20] Nie, J-Y. and Jin, F. A Multilingual Approach to Multilingual Information Retrieval. *Proceedings of the Cross-Language Evaluation Forum*, 2003, 101-110. Berlin: Springer-Verlag.

[21] Peters, C. (ed.). *Cross-Language Information Retrieval and Evaluation: Workshop of the Cross-Language Evaluation Forum, CLEF 2000*. Berlin: Springer-Verlag. 2001.

[22] Resnik, P., Olsen, M. B., and Diab, M. The Bible as a Parallel Corpus: Annotating the "Book of 2000 Tongues". *Computers and the Humanities*, 33, 1999, 129-153.

[23] Young, P. G. *Cross Language Information Retrieval Using Latent Semantic Indexing*. Master's thesis, University of Knoxville, Tennessee: Knoxville, TN, 1994.