# Evaluation of the Bible as a Resource for Cross-Language Information Retrieval

**Peter A. Chew** **Steve J. Verzi** **Travis L. Bauer** **Jonathan T. McClain**
Sandia National Laboratories
P. O. Box 5800
Albuquerque, NM 87185, USA
{pchew,sjverzi,tlbauer,jtmccl}@sandia.gov

## Abstract

An area of recent interest in cross-language information retrieval (CLIR) is the question of which parallel corpora might be best suited to tasks in CLIR, or even to what extent parallel corpora can be obtained or are necessary. One proposal, which in our opinion has been somewhat overlooked, is that the Bible holds a unique value as a multilingual corpus, being (among other things) widely available in a broad range of languages and having a high coverage of modern-day vocabulary. In this paper, we test empirically whether this claim is justified through a series of validation tests on various information retrieval tasks. Our results appear to indicate that our methodology may significantly outperform others recently proposed.

## 1 Introduction

This paper describes an empirical evaluation of the Bible as a resource for cross-language information retrieval (CLIR). The paper is organized as follows: section 2 describes the background to this project and explains our need for CLIR. Section 3 sets out the various alternatives available (as far as multilingual corpora are concerned) for the type of textual CLIR which we want to perform, and details in qualitative terms why the Bible would appear to be a good candidate. In section 4, we outline the mechanics behind the 'Rosetta-Stone' type method we use for cross-language comparison. The manner in which both this method, and the reliability of using the Bible as the basis for cross-language comparison, are validated is outlined in section 5, together with the results of our tests. Finally, we conclude on and discuss these results in section 6.

## 2 Background

This paper describes a project which is part of a larger, ongoing, undertaking, the goal of which is to harvest a representative sample of material from the internet and determine, on a very broad scale, the answers to such questions as:

- what ideas in the global public discourse enjoy most currency;
- how the popularity of ideas changes over time.

Ideas are, of course, expressed in words; or, to put it another way, a document's vocabulary is likely to reveal something about the author's ideology (Lakoff, 2002). In view of this, and since ultimately we are interested in clustering the documents harvested from the internet by their ideology (and we understand 'ideology' in the broadest possible sense), we approach the problem as a textual information retrieval (IR) task.

There is another level of complexity to the problem, however. The language of the internet is not, of course, confined to English; on the contrary, the representation of other languages is probably increasing (Hill and Hughes, 1998; Nunberg, 2000). Thus, for our results to be representative, we require a way to compare documents in one language to those in potentially any other language. Essentially, we would like to answer the question of how ideologically aligned two documents are, regardless of their respective languages. In cross-language IR, this must be approached by the use of a parallel multilingual corpus, or at least some kind of appropriate training material available in multiple languages.

## 3 Parallel multilingual corpora: available alternatives

One collection of multilingual corpora gathered with a specific view towards CLIR has been de-



*Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, pages 68–74,
Sydney, July 2006. 

veloped by the Cross-Language Evaluation Forum (CLEF); see, for example, Gonzalo (2001). This collection, and its most recent revision (at the CLEF website, www.clef-campaign.org), are based on news documents or governmental communications. Use of such corpora is widespread in much recent CLIR work; one such example is Nie, Simard, Isabelle and Durand (1999), which uses the Hansard corpus, parallel French-English texts of eight years of the Canadian parliamentary proceedings, to train a CLIR model.

It should be noted that the stated objective of CLEF is to 'develop and maintain an infrastructure for the testing and evaluation of information retrieval systems operating on European languages' (Peters 2001:1). Indeed, there is good reason for this: CLEF is an activity under the auspices of the European Commission. Likewise, the Canadian Hansard corpus covers only English and French, the most widespread languages of Canada. It is to be expected that governmental institutions would have most interest in promoting resources and research in the languages falling most within their respective domains.

But in many ways, not least for the computational linguistics community, nor for anyone interested in understanding trends in global opinion, this represents an inherent limitation. Since many of the languages of interest for our project are not European – Arabic is a good example – resources such as the CLEF collection will be insufficient by themselves. The output of global news organizations is a more promising avenue, because many such organizations make an effort to provide translations in a wide variety of languages. For example, the BBC news website (http://news.bbc.co.uk/) provides translations in 34 languages, as follows:

> Albanian, Arabic, Azeri, Bengali, Burmese, Chinese, Czech, English, French, Hausa, Hindi, Indonesian, Kinyarwanda, Kirundi, Kyrgyz, Macedonian, Nepali, Pashto, Persian, Portuguese, Romanian, Russian, Serbian, Sinhala, Slovene, Somali, Spanish, Swahili, Tamil, Turkish, Ukrainian, Urdu, Uzbek, Vietnamese

However, there is usually no assurance that a news article in one language will be translated into any, let alone all, of the other languages.

In view of this, even more promising still as a parallel corpus for our purposes is the Bible. Resnik, Olsen and Diab (1999) elaborate on some of the reasons for this: it is the world's most translated book, with translations in over 2,100 languages (often, multiple translations per language) and easy availability, often in electronic form and in the public domain; it covers a variety of literary styles including narrative, poetry, and correspondence; great care is taken over the translations; it has a standard structure which allows parallel alignment on a verse-by-verse basis; and, perhaps surprisingly, its vocabulary appears to have a high rate of coverage (as much as 85%) of modern-day language. Resnik, Olsen and Diab note that the Bible is small compared to many corpora currently used in computational linguistics research, but still falls within the range of acceptability based on the fact that other corpora of similar size are used; and as previously noted, the breadth of languages covered is simply not available elsewhere. This in itself makes the Bible attractive to us as a resource for our CLIR task. It is an open question whether, because of the Bible's content, relatively small size, or some other attribute, it can successfully be used for the type of CLIR we envisage. The rest of this paper describes our attempt to establish a definitive answer to this question.

## 4 Methods for Cross-Language Comparison

All of the work described in this section was implemented using the Sandia Text Analysis Extensible Library (STANLEY). STANLEY allows for information retrieval based on a standard vector model (Baeza-Yates and Ribeiro-Neto, 1999: 27-30) with term weighting based on log entropy. Previous work (Bauer et al 2005) has shown that the precision-recall curve for STANLEY is better than many other published algorithms; Dumais (1991) finds specifically that the precision-recall curve for information retrieval based on log-entropy weighting compares favorably to that for other weighting schemes. Two distinct methods for cross-language comparison are described in this section, and these are as follows.

The first method (Method 1) involves creating a separate textual model for each 'minimal unit' of each translation of the Bible. A 'minimal unit' could be as small as a verse (e.g. Genesis 1:1), but it could be a group of verses (e.g. Genesis 1:1-10); the key is that alignment is possible because of the chapter-and-verse structure of the Bible, and that whatever grouping is used should be the same in each translation. Thus, for each language λ we end up with a set of models ($m_{1,\lambda}$, $m_{2,\lambda}$, … $m_{n,\lambda}$). If the Bible is used as the parallel corpus and the 'minimal unit' is the verse, then $n$ = 31,102 (the number of verses in the Bible).

Let us suppose now that we wish to compare document $d_i$ with document $d_j$, and that we happen to know that $d_i$ is in English and $d_j$ is in Russian. In order to assess to what extent $d_i$ and $d_j$ are 'about' the same thing, we treat the text of each document as a query against all of the models in its respective language. So, $d_i$ is evaluated against $m_{1,English}$, $m_{2,English}$, …, $m_{n,English}$ to give $sim_{i,1}$, $sim_{i,2}$, …, $sim_{i,n}$, where $sim_{x,y}$ (a value between 0 and 1) represents the similarity of document $d_x$ in language λ to model $m_n$ in language λ, based on the cosine of the angle between the vector for $d_x$ and the vector for $m_n$. Similar evaluations are performed for $d_j$ against the set of models in Russian. Now, each set of $n$ results for a particular document can itself be thought of an $n$-dimensional vector. Thus, $d_i$ is associated with ($sim_{i,1}$, $sim_{i,2}$, …, $sim_{i,n}$) and $d_j$ with ($sim_{j,1}$, $sim_{j,2}$, …, $sim_{j,n}$). To quantify the similarity between $d_i$ and $d_j$, we now compute the cosine between these two vectors to yield a single measure, also a value between 0 and 1. In effect, we have used the multilingual corpus – the Bible, in this case – in 'Rosetta-Stone' fashion to bridge the language gap between $d_i$ and $d_j$. Method 1 is summarized graphically in Figure 1, for two hypothetical documents.

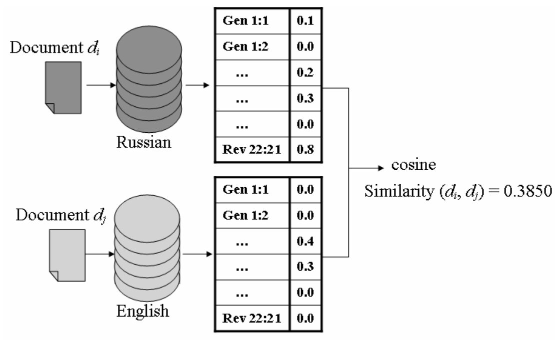

Figure 1: Method 1 for cross-language comparison

The second method of comparison (Method 2) is quite similar. This time, however, instead of building one set of textual models for each translation in language λ ($m_{1,\lambda}$, $m_{2,\lambda}$, … $m_{n,\lambda}$), we build a *single* set of textual models for *all* translations, with each language represented at least once ($m_1$, $m_2$, … $m_n$). Thus, $m_1$ might represent a model based on the concatenation of Genesis 1:1 in English, Russian, Arabic, and so on. In a fashion similar to that of Method 1, each incoming document $d_i$ is evaluated as a query against $m_1$, $m_2$, …, $m_n$, to give an $n$-dimensional vector where each cell is a value between 0 and 1. Method 2 is summarized graphically in Figure 2, for just English and Russian.

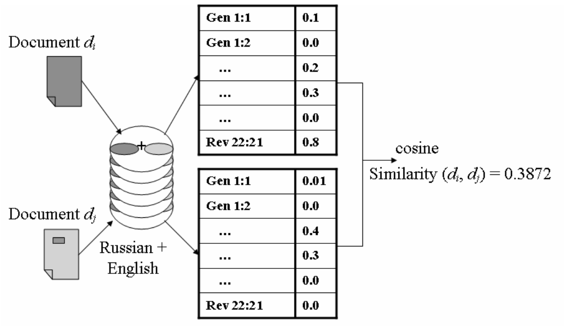

Figure 2: Method 2 for cross-language comparison

There are at least two features of Method 2 which make it attractive, from a linguist's point of view, for CLIR. The first is that it allows for the possibility that a single input document may be multilingual. In Figure 2, document $d_j$ is represented by an symbol with a mainly light-colored background, but with a small dark-colored section. This is intended to represent a document with mainly English content, but some small subsection in Russian. Under Method 1, in which $d_j$ is compared to an English-language model, the Russian content would have been effectively ignored, but under Method 2 this is no longer the case. Accordingly, the hypothetical similarity measure for the first 'minimal unit' has changed very slightly, as has the overall measure of similarity between document $d_i$ and $d_j$.

The second linguistic attraction of Method 2 is that it is not necessary to know a priori the language of $d_i$ or $d_j$, providing that the language is one of those for which we have textual data in the model set. Since, as already stated, the Bible covers over 2,100 languages, this should not be a significant theoretical impediment.

The theoretical advantages of Method 1 have principally to do with the ease of technical implementation. New model sets for additional languages can be easily added as they become available, whereas under Method 2 the entire model set must be rebuilt (statistics recomputed, etc.) each time a new language is added.

## 5 Validation of the Bible as a resource for CLIR

In previous sections, we have rehearsed some of the qualitative arguments for our choice of the

Bible as the basis for CLIR. In this section, we consider how this choice may be validated empirically. We would like to know how reliable the cross-language comparison methods outlined in the previous section are at identifying documents in different languages but which happen to be similar in content. This reliability will be in part a function of the particular text analysis model we employ, but it will also be a function of our choice of parallel text used to train the model. The Bible has some undeniable qualitative advantages for our purposes, but are the CLIR results based on it satisfactory in practice? Three tests are described in this section; the aim of these is to provide an answer to this question.

### 5.1 Preliminary analysis

In order to obtain a preliminary idea of whether this method was likely to work, we populated the entire matrix of similarity measures, verse by verse, for each language pair. There are 31,102 verses in the Bible (allowing for some variation in versification between different translations, which we carefully controlled for by adopting a common versification schema). Thus, this step involved building a 31,102 by 31,102 matrix for each language pair, in which the cell in row *m* and column *n* contains a number between 0 and 1 representing the similarity of verse *m* in one language to verse *n* in the other language. If use of the Bible for CLIR is a sound approach, we would expect to see the highest similarity measures in what we will call the matrix's diagonal values – the values occurring down the diagonal of the matrix from top-left to bottom-right – meaning that verse *n* in one language is most similar to verse *n* in the other, for all *n*.

Here, we would simply like to note an incidental finding. We found that for certain language pairs, the diagonal values were significantly higher than for other language pairs, as shown in Table 1.

| **Language pair** | **Mean similarity, verse by verse** |
|---|---|
| English-Russian | 0.3728 |
| English-Spanish | 0.5421 |
| English-French | 0.5508 |
| Spanish-French | 0.5691 |

**Table 1. Mean similarities by language pair**

One hypothesis we have is that the lower overall similarity for English-Russian is at least partly due to the fact that Russian is a much more highly inflected language then any of English, French, or Spanish. That many verses containing non-dictionary forms are the ones that score the highest for similarity, and many of those that do not score lowest, appears to confirm this. However, there appear to be other factors at play as well, since many of the highest-scoring verses contain proper names or other infrequently occurring lexical items (examples are Esther 9:9: 'and Parmashta, and Arisai, and Aridai, and Vaizatha', and Exodus 37:19: 'three cups made like almond-blossoms in one branch, a bud and a flower, and three cups made like almond-blossoms in the other branch, a bud and a flower: so for the six branches going out of the lampstand'). A third possibility, consistent with the first, is that Table 1 actually reflects more general measures of similarity between languages, the Western European languages (for example) all being more closely related to Latin than their Slavic counterparts. At any rate, if our hypothesis about inflection being an important factor is correct, then this would seem to underline the importance of stemming for highly-inflected languages.

### 5.2 Simple validation

In this test, the CLIR algorithm is trained on the entire Bible, and validation is performed against available extra-Biblical multilingual corpora such as the FQS (2006) and RALI (2006) corpora. This test, together with the tests already described, should provide a reliable measure of how well our CLIR model will work when applied to our target domain (documents collected from the internet).

For this test, five abstracts in the FQS (2006) were selected. These abstracts are in both Spanish and English, and the five are listed in Table 2 below.

| | |
|---|---|
| Eng. 1 | Perspectives |
| Eng. 2 | Public and Private Narratives |
| Eng. 3 | Qualitative Research |
| Eng. 4 | How Much Culture is Psychology Able to Deal With |
| Eng. 5 | Conference Report |
| Sp. 1 | Perspectivas |
| Sp. 2 | Narrativas públicas y privadas |
| Sp. 3 | Cuánta cultura es capaz de abordar la Psicología |
| Sp. 4 | Investigación cualitativa |
| Sp. 5 | Nota sobre la conferencia |

**Table 2. Documents selected for analysis**

The results based on these five abstracts, where comparison was performed between Spanish and English and vice-versa, are as shown in Table 3. The results shown in Table 3 are the actual (raw) similarity values provided by our CLIR framework using the FQS corpus.

| | Eng. 1 | Eng. 2 | Eng. 3 | Eng. 4 | Eng. 5 |
|---|---|---|---|---|---|
| **Sp. 1** | 0.6067 | 0.0430 | 0.0447 | 0.0821 | 0.1661 |
| **Sp. 2** | 0.0487 | 0.3969 | 0.0377 | 0.0346 | 0.0223 |
| **Sp. 3** | 0.1018 | 0.0956 | 0.0796 | 0.1887 | 0.1053 |
| **Sp. 4** | 0.0303 | 0.0502 | 0.0450 | 0.1013 | 0.0493 |
| **Sp. 5** | 0.0354 | 0.1314 | 0.0387 | 0.0425 | 0.1682 |

**Table 3. Raw similarity values of Spanish and English documents from FQS corpus**

In this table, 'Eng. 1', 'Sp. 1', etc., refer to the documents as listed in Table 2.

In four out of five cases, the CLIR engine correctly predicted which English document was related to which Spanish document, and in four out of five cases it also correctly predicted which Spanish document was related to which English document. We can relate these results to traditional IR measures such as precision-recall and mean average precision by using a query that returns the top-most similar document. Thus, our 'right' answer set as well as our CLIR answers will consist of a single document. For the FQS corpus, this represents a mean average precision (MAP) of 0.8 at a recall point of 1 (the first document recalled). The incorrect cases were Eng. 4, where Sp. 3 was predicted, and Sp. 3, where Eng. 4 was predicted. (By way of possible explanation, both these two documents included the keywords 'qualitative research' with the abstract.) Furthermore, in most of the cases where the prediction was correct, there is a clear margin between the score for the correct choice and the scores for the incorrect choices. This leads us to believe that our general approach to CLIR is at very least promising.

### 5.3 Validation on a larger test set

To address the question of whether the CLIR approach performs as well on larger test sets, where the possibility of an incorrect prediction is greater simply because there are more documents to select from, we trained the CLIR engine on the Bible and validated it against the 114 suras of the Quran, performing a four-by-four-way test using the original Arabic (AR) text plus English (EN), Russian (RU) and Spanish (ES) translations. The MAP at a recall point of 1 is shown for each language pair in Table 4.

| | | Language of predicted document | | | |
|---|---|---|---|---|---|
| | | **AR** | **EN** | **RU** | **ES** |
| **Language of input** | **AR** | 1.0000 | 0.2193 | 0.2281 | 0.2105 |
| | **EN** | 0.2632 | 1.0000 | 0.3333 | 0.5263 |
| | **RU** | 0.2719 | 0.3860 | 1.0000 | 0.4386 |
| | **ES** | 0.2105 | 0.4912 | 0.4035 | 1.0000 |

**Table 4. Results based on Quran test**

This table shows, for example, that for 52.63% (or 60) of the 114 English documents used as input, the correct Spanish document was retrieved first. As with the results in the previous section, we can relate these results to MAP at a recall of 1. If we were to consider more than just the top-most similar document in our CLIR output, we would expect the chance of seeing the correct document to increase. However, since in this experiment the number of relevant documents can never exceed 1, the precision will be diluted as more documents are retrieved (except at the point when the one correct document is retrieved). The values shown in the table are, of course, greater by a couple of orders of magnitude than that expected of random retrieval, of 0.0088 (1/114). Our methodology appears significantly to outperform that proposed by McNamee and Mayfield (2004), who report an MAP of 0.3539, and a precision of 0.4520 at a recall level of 10, for English-to-Spanish CLIR based on 5-gram tokenization. (We have not yet been able to compare our results to McNamee and Mayfield's using the same corpora that they use, but we intend to do this later. We do not expect our results to differ significantly from those we report above.) Perhaps not surprisingly, our results appear to be better for more closely-related languages, with pairs including Arabic being consistently those with the lowest average predictive precision across all suras.

## 6 Discussion

In this paper, we have presented a non-language-specific framework for cross-language information retrieval which appears promising at least for our purposes, and potentially for many others. It has the advantages of being easily extensible, and, with the results we have presented, it is empirically benchmarked. It is extensible in two dimensions; first, by language (substantially any human language which might be represented on the internet can be covered, and the cost of adding resources for each additional language is relatively small), secondly, by extending the training set with additional corpora, for available language pairs. Doubtless, also, the methodology could be further tuned for better performance.

It is perhaps surprising that the Bible has not been more widely used as a multilingual corpus by the computational linguistics and information retrieval community. In fact, it usually appears to be assumed by researchers that parallel texts, particularly those which have been as carefully translated as the Bible and are easy to align, are scarce and hard to come by (for two examples, see McNamee and Mayfield 2004 and Munteanu and Marcu 2006). The reason for the Bible being ignored may be the often unspoken assumption that the domain of the Bible is too limited (being a religious document) or that its content is too archaic. Yet, the truth is that much of the Bible's content has to do with enduring human concerns (life, death, war, love, etc.), and if the language is archaic, that may have more a matter of translation style than of content.

There are a number of future research directions in computational linguistics we would like to pursue, besides those which may be of interest in other disciplines. The first is to use this framework to evaluate the relative faithfulness of different translations. For example, we would expect to see similar statistical relationships within the model for a translation of the Bible as are seen in its original languages (Hebrew and Greek). Statistical comparisons could thus be used as the basis for evaluating a translation's faithfulness to the original. Such an analysis could be of theological, as well as linguistic, interest.

Secondly, we would like to examine whether the model's performance can be improved by introducing more sophisticated morphological analysis, so that the units of analysis are morphemes instead of words, or possibly morphemes as well as words.

Third, we intend to investigate further which of the two methods outlined in section 4 performs better in cross-language comparison, particularly when the language of the source document is unknown. In particular, we are interested in the extent to which homographic cognates across languages (e.g. French *coin* 'corner' versus English *coin*), may affect the performance of the CLIR engine.

## Acknowledgement

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.

## References

Lars Asker. 2004. Building Resources: Experiences from Amharic Cross Language Information Retrieval. Paper presented at *Cross-Language Information Retrieval and Evaluation: Workshop of the Cross-Language Evaluation Forum, CLEF 2004.*

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. New York: ACM Press.

Travis Bauer, Steve Verzi, and Justin Basilico. 2005. Automated Context Modeling through Text Analysis. Paper presented at *Cognitive Systems: Human Cognitive Models in System Design*.

Susan Dumais. 1991. Improving the Retrieval of Information from External Sources. *Behavior Research Methods, Instruments, and Computers* 23(2):229-236.

Forum: Qualitative Social research (FQS). 2006. *Published Conference Reports.* (Conference reports available on-line in multiple languages.) Accessed at http://www.qualitative-research.net/fqs/conferences/conferences-pub-e.htm on February 22, 2006.

Julio Gonzalo. 2001. Language Resources in Cross-Language Text Retrieval: a CLEF Perspective. In Carol Peters (ed.). *Cross-Language Information Retrieval and Evaluation: Workshop of the Cross-Language Evaluation Forum, CLEF 2000*: 36-47. Berlin: Springer-Verlag.

George Lakoff. 2002. *Moral politics : how liberals and conservatives think.* Chicago : University of Chicago Press.

Paul McNamee and James Mayfield. 2004. Character *N*-Gram Tokenization for European Language Text Retrieval. *Information Retrieval* 7: 73-97.

Dragos Munteanu and Daniel Marcu. 2006. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics* 31(4):477-504.

Geoffrey Nunberg. 2000. Will the Internet Always Speak English? *The American Prospect* 11(10).

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

Recherche appliquée en linguistique informatique (RALI). 2006. *Corpus aligné bilingue anglais-français*. Accessed at http://rali.iro.umontreal.ca/ on February 22, 2006.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The Bible as a Parallel Corpus: Annotating the "Book of 2000 Tongues". *Computers and the Humanities*, 33: 129-153.