Evaluating Cross-Language Text Filtering Effectiveness *

Douglas W. Oard, College of Library and Information Services
Bonnie J. Dorr, Department of Computer Science
University of Maryland, College Park, MD 20742
{oard|bonnie}@umiacs.umd.edu

Abstract

In this paper we describe an evaluation methodology for cross-language text filtering systems which exploits existing test collections that were designed for monolingual evaluations. Our methodology, based on normative relevance assessments by expert users, is well suited for comparing the effect of different cross-language mapping techniques on filtering accuracy. By measuring the degradation introduced by the use of existing test collections, we are able to qualify the broader applicability of our results and to quantify the improvement in evaluation accuracy that would result from development of a test collection tailored to the evaluation of multilingual text filtering systems. From our experiments we conclude that the additional investment required to produce a truly multilingual test collection would be well justified because evaluation of multilingual text filtering techniques appears to be both practical and productive.

1 Introduction

Text filtering is a special case of the document detection problem in which we seek to model relatively stable interests and then use that model to improve the way in which information about document availability is presented to the user. For example, a text filtering system might observe the user's reading behavior and then construct a sorted list of newly arrived documents in which the documents at the head of the list are those predicted to be most likely to be read. Our goal is to construct a multilingual text filtering system in which documents in many languages can be handled. In this paper we describe an evaluation methodology for such systems which exploits existing test collections.

Dramatic reductions in the cost of communicating, storing and processing information have produced an explosion of international communications, much of it either in text form or with text annotations. In the emerging worldwide information economy, text filtering systems capable of managing documents in multiple languages would have obvious utility in financial, diplomatic, news, academic, and even entertainment applications. Surprisingly, we are not aware of any prior research in which evidence obtained by observing reading behavior for documents in one language is used to enhance the predictions of future user behavior on documents in another language, a problem we call "adaptive multilingual text filtering."

Two established research areas together offer a range of useful approaches for multilingual text filtering applications. Multilingual text retrieval systems seek to select documents in one language based on queries expressed in another, and some monolingual text filtering systems seek to learn information need representations based on observations of user behavior. The focus of our research has been to explore how these approaches can be exploited to satisfy the unique requirements of adaptive multilingual text filtering.

The cross-language training aspect of adaptive multilingual text filtering introduces unique challenges for performance evaluation. The methodology we have adopted, based on normative relevance assessments by expert users, is well suited for comparing the effect of different approaches on prediction accuracy. Because it requires large document collections and expensive manually determined relevance assessments, we have developed an application of the method which exploits existing test collections. By measuring the degradation introduced by the characteristics of those collections, we are able to qualify the broader applicability of our

*This work has been supported in part by NSF award IRI-9357731, ARPA and ONR contract N00014-92-J-1929, ARPA contract DACA76-92-C009, and the Logos Corporation.

results and to quantify the improvement in evaluation accuracy that would result from development of a test collection tailored to the evaluation of multilingual text filtering systems.

2 Adaptive Multilingual Text Filtering

We have surveyed text filtering techniques elsewhere [1], so here we describe only the technique which we have chosen to apply. Our approach is based on the ranked output paradigm in which the text filtering system seeks to rank order newly arrived documents with the most useful documents near the top of the list. We have based our work on a technique developed by Dumais for monolingual text filtering in which Latent Semantic Indexing (LSI) is used to develop relatively short feature vectors that describe the relevant training documents, and the mean of the relevant documents' feature vectors is used as the "profile" (information need representation) [2]. LSI feature vectors describing newly arrived documents are then used to rank order the newly arrived documents in order of decreasing similarity with the profile. We use the cosine similarity measure because it emphasizes content similarity while suppressing the effect of document length variations.

LSI feature vectors are constructed by counting the frequency with which each term occurs in a document and then using those values as input to a function which reduces the number of features by accounting for similarities in word usage. This function is automatically constructed using statistical techniques by examining a representative collection of text in which typical term usage variations are exhibited. We have applied this "LSI-mean" filtering approach to evaluate the performance of three cross-language mapping techniques, so we have been careful to construct this mapping using the same document collection in order to assure the comparability of our results.

Two of the cross-language mapping techniques we are evaluating are motivated by earlier work on multilingual text retrieval, a topic we have also surveyed [3]. The most obvious is to pass every document through an automatic machine translation system. In multilingual text retrieval it is the "query" (the information need specification) which is most often translated. While the brevity of typical queries makes that choice efficient, use of machine translation with the LSI-mean text filtering technique requires that every document be translated into a single language because the LSI-mean profile is a vector made up of elements which do not correspond to individual words. Our approach, which we call "Text Translation," effectively reduces multilingual text filtering to it's monolingual counterpart.

A second technique, which we call "Latent Semantic Coindexing," exploits the ability of LSI to identify and suppress the effect of word usage variations. In Latent Semantic Coindexing, bilingual or multilingual documents are prepared by adjoining versions of the same document in different languages. LSI is then trained on that document collection to find a feature vector mapping which accepts documents from any of the languages [4]. It is our interest in this technique which led us to choose the LSI-mean technique as our standard text filtering method.

Other approaches to multilingual text filtering are possible as well, and we have used the same methodology to evaluate a third technique which we call Vector Translation. We limit our discussion here to Text Translation and Latent Semantic Coindexing since two techniques suffice to illustrate our evaluation methodology.

3 Ideal Experiment Design

Because we wish to characterize the effect of employing three different cross-language mapping techniques, we concentrate on the quality of the rank ordering produced by the LSI-based text filtering technique. The quality of a rank ordering is often evaluated using "precision" at one or more values of "recall" and that is the technique we have adopted [5]. Using a standard document collection and topic set, each document is evaluated by human experts to determine whether it is relevant to each topic.[1] A set of documents is then chosen beginning at the top of the ranked list and proceeding as far down as necessary to achieve the desired level of comprehensiveness ("recall"—the fraction of the relevant documents which are included in the set), and the concentration of relevant documents in that set ("precision"—the fraction of the set that is relevant) is then computed. Precision is often averaged over several values of recall to compute a single figure of merit

[1] For large test collections a sampling technique is often used to limit the cost of generating relevance judgements.

Partition	English	Spanish	Relevance Judgements
Cross-Language Training	X	X	
Profile Training	X		X
Effectiveness Evaluation		X	X

Table 1: Ideal multilingual text filtering test collection.

for a topic. We have chosen instead to report precision only at a fixed value of recall (0.1—the point at which 10% of the relevant documents have been seen.) The density of relevant documents is greatest near the top of the ranked list, so differences in cross-language mapping effectiveness should be most apparent at in that region. In our experiments, a recall of 0.1 is achieved after 35, 36 or 8 documents (for topics SP22, SP25 and SP47 respectively) have been found. Since that should be an adequate number of relevant documents for many types of interactive applications, we believe that the precision values we report are representative of what might be experienced by interactive users.

Text filtering experiments of the type we are conducting require a document collection for which relevance judgements are available, so it would be ideal to construct a test collection in which every document has, for example, both English and Spanish versions, as well as relevance judgements with respect to a number of standardized topics. While we ultimately intend to provide users with systems which adapt in nearly real time, for our evaluation we have chosen to introduce an artificial division between the construction of a profile and the use of that profile to rank order documents. We could achieve this by dividing an ideal test collection into two partitions, one for profile training and one for effectiveness evaluation. Because we wish to measure the effectiveness of cross-language selection, we use the documents in English from one partition and their associated relevance judgements to develop the profile. We then apply a cross-language ranking system to rank order the Spanish documents from the other partition, using their associated relevance judgements to determine the quality of that ranking. We have chosen English for profile training and Spanish for evaluation because that choice simplified the design of our Text Translation experiment. We used the same selections for the Latent Semantic Coindexing experiment in order to obtain comparable results.

In Latent Semantic Coindexing we seek to extract statistical information about word cooccurrence from a large collection of documents in which every document is duplicated in each language. In order to apply that technique we would need to select a third partition of the test collection from which we can extract collocation information. Cross-language text selection would not be needed if the documents in the profile training and the evaluation partitions were available in both languages, so it would not be reasonable to reuse one of the existing partitions for this "cross-language training" task. Relevance judgements are not used for language training. Table 1 shows which parts of the three partitions of an ideal test collection would be used.

4 Use of Available Corpora

We are aware of no large collection of the type shown in Figure 1. Large bilingual and trilingual document collections exist, but construction of the required topics and relevance judgements would be a massive undertaking. Large monolingual collections with topics and relevance judgements also exist, but translation of each document into a second language would be even less feasible. Because no partition in Figure 2 requires both aspects (bilingual and scored), it would be possible to reduce the expense somewhat by constructing each portion of the evaluation collection independently. We have taken this concept one step further and identified three existing document collections which can be used together to approximate the results that would be achieved using an ideal test collection. The collections we have used are shown in Table 2. All three collections are available through the Linguistic Data Consortium[2]. The topics and relevance judgements for the Wall Street Journal and El Norte (a Mexican newspaper) collections were developed using a pooled

[2]http://ftp.cis.upenn.edu/pub/ldc_www/hpage.html

Source	English	Spanish	English Rel.	Spanish Rel.
1990-1992 UN Documents	X	X		
1990-1992 Wall St Journal	X		X	
1992 El Norte Newspaper		X		X

Table 2: Evaluation using existing collections.

Spanish Language Topic		English Language Topic	
SP10	Mexican Narcotic Trafficking	284	International Drug Enforcement
SP18	Foreign Car Makers in Mexico	290	Foreign Car Makers in the U.S.
SP22	Mexican Inflation	008	Economic Projections
SP25	Mexican Privatization Programs	128	Privatization of State Assets
SP47	Mexican Cancer Cause Research	123	Carcinogen Research and Control

Table 3: Closely related English and Spanish TREC topics.

relevance assessment methodology as part of the U.S. National Institute of Standards and Technology's (NIST) Text REtrieval Conferences (TREC).

Two potential problems arise when the three existing collections in Figure 3 are substituted for the single collection shown in Figure 1. The first is that the domains addressed by the UN, the Wall Street Journal and El Norte would be expected to differ significantly. We refer to this problem as a "domain shift." A potentially even more serious problem is that the Wall Street Journal and El Norte articles were not judged against the same topics. The "English Rel." relevance judgements identify the relevance of the Wall Street Journal articles to 250 topics, while the "Spanish Rel." judgements specify the relevance of the El Norte articles to 50 independently chosen topics. We call this problem "topic shift."

Table 3 shows the five Spanish topics for which we have found closely corresponding English topics. Although the detailed topic descriptions that are distributed with the collections identify some differences, there is sufficient overlap to suggest that a minimal adjustment to the sets of relevant documents would result in comparable sets of documents in the two languages. In fact, our experimental results confirm that it is possible to use the relevance judgements without any adjustment when the goal is to compare different cross-language mapping techniques.

The domain shift between the UN documents and one of the newspapers (El Norte) is fairly easy to evaluate. In order to ensure that we obtain comparable results, we have chosen to use the LSI-mean filtering technique for Text Translation and Latent Semantic Coindexing. Since Text Translation produces Spanish documents as an intermediate step, we can measure the effect of the domain shift by running the Text Translation experiment a second time. In that second run we substitute the El Norte documents for the Spanish UN documents when generating the mapping that produces the LSI feature vectors. The resulting LSI mapping will be better suited to the El Norte articles, and the difference in our precision measure reveals the effect of the domain shift between the UN collection and the El Norte collection. We have not developed any similar technique to reveal the effect of the topic shift between either of those collections and the Wall Street Journal collection.

We can estimate the effect of the topic shift by comparing cross-language and within-language performance. This can be done by dividing the El Norte collection into two partitions and then performing a monolingual evaluation in which one partition is used for profile training and the other for evaluation. This removes the effect of the topic shift completely, although it simultaneously removes the effect of errors introduced by the cross-language mapping technique. The effect of translation errors on the performance of the Text Translation technique are easily measured, however, using a modification of the basic Latent Semantic Coindexing experiment. With Latent Semantic Coindexing, LSI feature vectors can be produced

Topic Pair	Technique		
	LSC	TT	None
SP22/008	0.17	0.17	0.06
SP25/128	0.08	0.10	0.03
SP47/123	0.07	0.06	0.00

Table 4: Multilingual text filtering experiment results (precision at 0.1 recall).

from either English or Spanish documents. If the English Wall Street Journal articles are translated into Spanish before being used for profile training in the Latent Semantic Coindexing experiment, the observed reduction in precision will be entirely attributable to errors introduced by the machine translation step. These are exactly the same errors that affect the Text Translation experiment, so this result will reveal the necessary adjustment to the difference between the monolingual evaluation on El Norte and the standard Text Translation experiment. We have not yet conducted this experiment, but preliminary results in which we used the entire El Norte collection for both training and evaluation are reported below. Those results overstate the effect of the topic shift because they evaluate memory, not prediction accuracy, but they do provide an upper bound on the magnitude of the topic shift.

5 Results

TREC relevance judgements for topics 284 and 290 will not be available from NIST until October 1996, so we have only been able to use the last three topic pairs shown in Table 3. Table 4 shows results for two cross-language text filtering techniques, Latent Semantic Coindexing (LSC) and Text Translation (TT), and a baseline run (labeled "None") in which we used no cross-language mapping technique at all. These results are described in detail in [6]. In this paper we will limit our comments to those which address fundamental evaluation issues.

The most significant observation that we can draw from our experiments is that multilingual text filtering is practical and that the presently available corpora are adequate to demonstrate that fact. Both corpus-based techniques (such as Latent Semantic Coindexing) and knowledge-based techniques (such as Text Translation) have demonstrated better performance than that which could be achieved with no translation component, despite the limitations imposed by the topic and domain shifts. This fact should be of interest to researchers working on corpus-based multilingual text retrieval as well, since it confirms that (for these three topics, at least), the UN collection and the El Norte collection are sufficiently similar to produce much better precision near the top of the ranked list than that which could be achieved by random selection. In every case the precision achieved by random selection would have been below 0.01 at any value of recall. Additional details on this point are presented in [6].

Another interesting observation is that the results without cross-language mapping exhibit a surprising amount of variation. We attribute this effect to the existence of words which are common to Spanish and English that are useful for recognizing documents that are relevant to some topics. This observation has led us to conclude that when the available corpora limit a cross-language filtering or retrieval experiment to a small number of topics, a baseline run with no cross-language mapping is a simple way to gain some useful insight into the significance of the results.

Table 5 shows the results of the domain shift experiment. In two cases out of three, the domain shift between the UN collection and the El Norte collection appears to be substantial but not overwhelming. The lack of a clear domain shift effect in the third case is at least partially explained by poor performance of the LSI-mean filtering technique on topic SP25. In a completely monolingual evaluation memory (LSI training, profile training and evaluation all using the complete El Norte collection), the precision achieved by the LSI-mean technique at 0.1 recall was only 0.18. This poor performance could result from a number of factors (e.g., we used less than 2% of the available documents when El Norte was used for LSI training and those documents may have been poorly chosen), and we have not yet completed our evaluation of the cause of this

Topic Pair	LSI training	
	Spanish UN	El Norte
SP22/008	0.17	0.28
SP25/128	0.10	0.10
SP47/123	0.06	0.17

Table 5: Domain shift results for Text Translation (precision at 0.1 recall).

Topic Pair	Experiment Design		LSC profile training	
	Multilingual	Monolingual	English WSJ	Translated WSJ
SP10/022	0.02	0.20	0.01	0.01
SP22/008	0.17	0.46	0.17	0.14
SP25/128	0.10	0.10	0.08	0.13
SP47/123	0.06	0.45	0.07	0.02

Table 6: Preliminary topic shift results (precision at 0.1 recall).

deficiency.

Table 6 shows preliminary results which provide bounds on the magnitude of the topic shift effect. Results for a fourth topic pair which we tried, SP10/022, are shown as well in order to illustrate the topic shift effect clearly. It appeared from inspection of the topic descriptions that topics SP10 and 022 were as similar as any of the other pairs we had chosen, but these results clearly reveal that that topic pair is not useful. Again, the SP25/128 topic pair yields unusual and as yet unexplained results, actually increasing precision when translation errors are introduced. The remaining two topic pairs show relatively large topic shift effects (although these are only upper bounds) after considering the relatively small translation error effects.

6 Conclusions

We have developed a way to apply existing collections to compare the effectiveness of cross-language mapping techniques in an adaptive multilingual text filtering system. The domain shift effect will be unavoidable for corpus-based techniques such as Latent Semantic Coindexing when the available collections of translated texts do not use language in exactly the same way as the newly arriving documents that must be filtered. Thus, the ability to characterize the magnitude of the domain shift effect will be important whenever knowledge-based and corpus-based techniques are compared. The topic shift effect, on the other hand, is strictly an artifact of our experiment design. Although we are able to estimate (or at least bound) the effect of the topic shift, it would clearly be better if a test collection were available with relevance judgements for documents in several languages with respect to an identical set of topics. The ongoing TREC evaluation provides an excellent venue for such an effort, since a set of relevance judgements on a multilingual document collection would facilitate monolingual evaluations in multiple languages as well as cross-language retrieval and filtering evaluations. The large collection on United Nations documents is available in three languages, making it an excellent candidate for this purpose. It is not possible to draw broadly applicable conclusions from only three topic pairs, but our results do at least indicate that the additional investment required to produce a truly multilingual test collection would be well justified because evaluation of adaptive multilingual text filtering techniques appears to be both practical and productive.

Acknowledgements

The authors would like to express their appreciation to David Hull of Rank Xerox Research Centre for his assistance with data preparation and to the Logos corporation for machine translation support.

References

[1] Douglas W. Oard and Gary Marchionini, "A conceptual framework for text filtering," Tech. Rep. CS-TR-3643, University of Maryland, May 1996, http://www.ee.umd.edu/medlab/filter/papers/filter.ps.

[2] S. T. Dumais, "Latent semantic indexing (LSI): TREC-3 report," in *Overview of the Third Text REtrieval Conference*, Donna Harman, Ed. Nov. 1994, pp. 219–230, NIST, http://potomac.ncsl.nist.gov/TREC/.

[3] Douglas W. Oard and Bonnie J. Dorr, "A survey of multilingual text retrieval," Tech. Rep. UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies, Apr. 1996, http://www.ee.umd.edu/medlab/filter/papers/mlir.ps.

[4] Thomas K. Landauer and Michael L. Littman, "Fully automatic cross-language document retrieval using latent semantic indexing," in *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, pp. 31–38. UW Centre for the New OED and Text Research, Waterloo Ontario, October 1990, http://www.cs.duke.edu/~mlittman/docs/x-lang.ps.

[5] Gerard Salton and Michael J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.

[6] Douglas William Oard, *Adaptive Vector Space Text Filtering for Monolingual and Cross-Language Applications*, Ph.D. thesis, University of Maryland, College Park, Aug. 1996, To appear. http://www.ee.umd.edu/medlab/filter/papers/thesis.ps.