Alignment of Spanish and English TREC Topic Descriptions *

Douglas W. Oard
College of Library and Information Services
University of Maryland, College Park, MD 20742
oard@glue.umd.edu

Abstract

A technique is described for aligning TREC topic descriptions that is capable of producing a small multilingual test collection which can be used for cross-language ad-hoc and routing evaluations. Methods for measuring the degree of degradation induced by the necessary approximations are described and illustrated using examples from an evaluation of two cross-language routing techniques. Although the experiments were conducted on a relatively small test collection using existing TREC relevance judgments, the results suggest that cross-language routing is practical and that the investment required to produce a cross-language test collection for the TREC multilingual track would be justified.

1 Introduction

The principal goal of the University of Maryland's participation in the Fifth Text REtrieval Conference (TREC-5) was to evaluate the performance of advanced routing techniques. We investigated two aspects of the routing problem: construction of a feature space that is independent of the language in which a document is written, and efficient construction of reduced-dimensional feature spaces on which machine learning algorithms can be effective. Our ultimate goal is to construct an adaptive multilingual routing system which can learn (from the user's responses to existing documents) to select new documents that may not be written in the same language as the existing documents. Our work on efficient construction of reduced-dimensional feature spaces revealed that the technique we applied was inappropriate for large document collections, so our official ad-hoc retrieval run was submitted using a version of SMART that was modified locally for the ISO 8859-1 character set.[1] Additional details regarding our development of monolingual routing techniques are reported in [7].[2]

In this paper we describe how we have used TREC document collections and relevance judgments to evaluate the performance of two techniques that we implemented for for cross-language routing because we believe that those techniques may be of use to other TREC participants. The multilingual routing systems we are exploring are based on existing approaches to ad-hoc cross-language text retrieval which seek to select documents in one language based on queries expressed in another, and on existing monolingual routing research. Our contribution has been to explore how these two bodies of research can best be exploited to satisfy the unique requirements of adaptive multilingual routing. Evaluation of the resulting systems has been our greatest challenge. The test collection we have constructed can be used to compare the effect of different cross-language mapping techniques on prediction accuracy, and we have developed a methodology for measuring the degradation introduced by the unavoidable compromises that we have made when constructing the collection. As a result, 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 routing systems.

---

*This work has been supported in part by DOD contract MDA9043C7217, ARPA and ONR contract N00014-92-J-1929, ARPA contract DACA76-92-C009, NSF award IRI-9357731, and the Logos Corporation.

[1] We used ltc term weights and participated in category B.

[2] Current links to most of the references cited in this paper can be found online at http://www.ee.umd.edu/medlab/filter/

2 Adaptive Multilingual Routing

We have surveyed text routing techniques elsewhere [6], 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 routing 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 routing 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 routing query [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 routing query using the cosine similarity measure.

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" routing approach to evaluate the performance of two 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.

The cross-language mapping techniques we have evaluated were motivated by earlier work on multilingual text retrieval, a topic we have also surveyed [8]. The most obvious is to pass every document through an automatic machine translation system. In ad-hoc cross-language retrieval it is the topic specification which is most often translated. While the brevity of typical topic specifications makes that choice efficient, use of machine translation with the LSI-mean routing technique requires that every document be translated into a single language because the LSI-mean routing query is a vector made up of elements which do not correspond to individual words. Our approach, which we call "Text Translation," effectively reduces cross-language routing to it's monolingual counterpart.

A second technique, "Cross-Language Latent Semantic Indexing," exploits the ability of LSI to identify and suppress the effect of word usage variations. In Cross-Language Latent Semantic Indexing, 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 [3, 4]. It is our interest in this technique which led us to choose the LSI-mean technique as the routing method.

Other approaches to multilingual routing 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 Cross-Language Latent Semantic Indexing since two techniques suffice to illustrate the way in which we have used TREC collections and relevance judgments and our work with Vector Translation is still quite preliminary.

3 Ideal Experiment Design

Routing experiments of the type we are conducting require a document collection for which relevance judgments are available, so it would be ideal if a test collection existed in which every document has versions in two languages and relevance judgments 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 routing query and the use of that routing query to rank order documents. We could achieve this by dividing an ideal test collection into two partitions, one for query construction 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 judgments to develop the routing query. We then apply a cross-language ranking system to rank order the Spanish documents from the other partition, using their associated relevance judgments to determine the quality of that ranking. We have chosen English for query construction and Spanish for evaluation because the Logos machine translation system we used for the Text Translation experiments was capable of unidirectional English to Spanish translation. We used the same selections for the Cross-Language Latent Semantic Indexing experiment in order to obtain comparable results.

Partition	English	Spanish	Relevance Judgments
Cross-Language Training	X	X	
Query Construction	X		X
Effectiveness Evaluation		X	X

Table 1: Ideal multilingual routing test collection.

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.

In Cross-Language Latent Semantic Indexing 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 need to select a third partition of the test collection from which we can extract collocation information. It would not be reasonable to reuse one of the existing partitions for this "cross-language training" task because cross-language techniques would not be needed if all of the documents in either the query construction or the evaluation partition were already available in both languages. Relevance judgments are not needed for this language training partition. Table 1 shows which parts of the three partitions of an ideal test collection are needed.

4 Use of TREC Collections

We are aware of no large collection of the type shown in Figure 1 (but for a description of a smaller collection in Korean and English see [5]), and large collections are needed for evaluation of adaptive routing systems. Large bilingual and trilingual document collections do exist, but construction of the required topics and relevance judgments would have been well beyond our resources. TREC has provided large monolingual collections with associated topics and relevance judgments, but translation of each document into a second language would have been even more difficult. Because none of the three partitions shown in Table 1 must be both bilingual and scored, it is possible to use three existing collections to approximate the results that would be achieved using an ideal test collection. The collections we have chosen are shown in Table 2. The UN collection is a large collection of United Nations documents, each of which is is available in either two or three languages (English, Spanish and French) from the Linguistic Data Consortium[3]. This is the same collection that was used for cross-language ad-hoc retrieval experiments by the New Mexico State University team in TREC-4 [1]. The topics and relevance judgments for the Wall Street Journal and El Norte collections are obtained from TIPSTER disk 2 and from the TREC-4 multilingual track respectively. We chose the Wall Street Journal collection on disk 2 because relevance judgments are available for that collection on all 300 topics and because that collection contains material from the same time period as the El Norte collection. For consistency, we used only those UN documents that were prepared during the same years as the Wall Street Journal articles that we used.

We performed topic alignment manually, examining each of the 50 Spanish topics and then scanning a list of the 300 available English topics in order to identify possible matches. The detailed topic descriptions were then compared and a set of topic pairs which appeared to be closely aligned were selected. Table 3 shows the five Spanish topics for which we have found closely corresponding English topics.[4] Although the

[3]Information on the availability of the UN collection can be obtained from http://www.ldc.upenn.edu

[4]Some more weakly aligned topic pairs that might also be useful are identified in [9].

Abbreviated Spanish Language Topic		Abbreviated 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.

topic descriptions in each pair have some differences, there is sufficient apparent 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 judgments without any adjustment when the goal is to compare different cross-language mapping techniques.

Two potential problems arise when the three existing collections in Table 2 are substituted for the single collection shown in Table 1. The first is that the subjects 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," between the collections since it is caused by differences in the topical domains of the two collections. A potentially even more serious problem is that the Wall Street Journal and El Norte articles were judged against topics which have been aligned after the fact, and that alignment is far from perfect. We call this problem "topic shift."

The domain shift between the UN documents and El Norte is fairly easy to evaluate. In order to ensure that we obtain comparable results, we have chosen to use the LSI-mean routing technique for Text Translation and Cross-Language Latent Semantic Indexing. 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 could be done by dividing the El Norte collection into two partitions and then performing a monolingual evaluation in which one partition is used for query construction and the other for evaluation. That would remove the effect of the topic shift completely, although it would simultaneously remove 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 Cross-Language Latent Semantic Indexing experiment. With Cross-Language Latent Semantic Indexing, LSI feature vectors can be produced from either English or Spanish documents. If the English Wall Street Journal articles are translated into Spanish before being used for query construction in the Cross-Language Latent Semantic Indexing 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. In our initial experiments we have used the entire El Norte collection for both training and evaluation when evaluating the topic shift. 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, and that upper bound proved to be adequate to recognize one case in which an extreme topic shift made an apparently well-aligned topic pair unusable.

Topic Pair	Technique		
	CL-LSI	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 routing experiment results (precision at 0.1 recall).

5 Results

Our primary objective is to determine the *relative* performance of two cross-language routing techniques. We would expect to find the largest absolute differences in precision near the top of the ranked list, and hence we felt that vales of precision at low recall would best reveal differences in performance between the two cross-language routing techniques that we tried. Thus, rather than report average precision, we have chosen 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, the precision values we report should be representative of what might be experienced by interactive users.

We used the SMART text retrieval system, modified locally to include the LSI-mean routing technique, with ltc term weights for our experiments. We substituted morphological roots provided by the Rank Xerox morphological tagger for SMART stemming because a third technique that we are developing (Vector Translation) could potentially benefit from compatibility with a bilingual dictionary. Relevance judgments for topics 284 and 290 were not available when we ran our experiments, so we were only able to use the last three topic pairs that are shown in Table 3. Table 4 shows results for the two cross-language routing techniques, Cross-Language Latent Semantic Indexing (CL-LSI) 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 [9]. In this paper we will limit our comments to those which address fundamental evaluation issues.

The most significant observation that we drew from our experiments is that multilingual routing appears to be practical and that the corpora we used are adequate to demonstrate that. Both corpus-based techniques (such as Cross-Language Latent Semantic Indexing) 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 also be of interest to researchers working on corpus-based ad-hoc cross-language retrieval, 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 [9].

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 routing 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 routing technique on topic SP25. In a completely monolingual evaluation of "memory" (LSI training, query construction and evaluation all using the complete El Norte collection), the precision

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		CL-LSI Query Construction	
	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).

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 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 SP10/022 is not a useful topic pair. 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 Future Work

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 judgments for documents in several languages with respect to an identical set of topics. The TREC-6 cross-language track will provide an initial step in this direction, assembling a document collection that could be used in future years for such an evaluation. One possible approach to exploiting this collection would be to design English topics in the main ad hoc retrieval task for which a suitable number of relevant documents would also be found in the cross-language collections. Relevance judgments generated in the ad hoc task could then be used in future years to build routing queries for the cross-language track. Although this would make topic selection for the ad hoc task more complex, it has the advantage of creating a one-direction (English to another language) cross-language routing evaluation collection without the expense of implementing pooled relevance assessments for the full cross-language track.

7 Conclusions

We have developed a way to apply existing TREC collections to compare the effectiveness of cross-language mapping techniques in an adaptive multilingual routing system. The domain shift effect we have described will be inherent in corpus-based techniques such as Cross-Language Latent Semantic Indexing, unless collections of translated texts which use language in almost exactly the same way as the training and evaluation documents can be found. 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. 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 cross-language test collection would be well justified because evaluation of adaptive multilingual routing techniques appears to be both practical and productive.

Acknowledgments

The author would like to express his appreciation to Bonnie Dorr and Michael Littman for their comments on earlier draft of this work, 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] Mark Davis and Ted Dunning. A TREC evaluation of query translation methods for multi-lingual text retrieval. In D. K. Harman, editor, *The Fourth Text Retrieval Conference (TREC-4)*. NIST, November 1995.

[2] S. T. Dumais. Latent Semantic Indexing (LSI): TREC-3 report. In Donna Harman, editor, *Overview of the Third Text REtrieval Conference*, pages 219–230. NIST, November 1994.

[3] Susan T. Dumais, Thomas K. Landauer, and Michael L. Littman. Automatic cross-linguistic information retrieval using latent semantic indexing. In Gregory Grefenstette, editor, *Working Notes of the Workshop on Cross-Linguistic Information Retrieval*. ACM SIGIR, August 1996.

[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*, pages 31–38. UW Centre for the New OED and Text Research, Waterloo Ontario, October 1990.

[5] Joon Ho Lee and Jeong Soo Ahn. Using n-grams for Korean text retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, August 1996.

[6] Douglas W. Oard. The state of the art in text filtering. *User Modeling and User-Adapted Interaction*, 1997. To appear.

[7] Douglas W. Oard and Nicholas DeClaris. Cognitive models for text filtering. Technical Report EE-TR-96-28, University of Maryland, College Park, 1996.

[8] Douglas W. Oard and Bonnie J. Dorr. A survey of multilingual text retrieval. Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies, April 1996.

[9] Douglas William Oard. *Adaptive Vector Space Text Filtering for Monolingual and Cross-Language Applications*. PhD thesis, University of Maryland, College Park, August 1996.