Resource Selection for
Domain-Specific Cross-Lingual IR

Monica Rogati   Yiming Yang

School of Computer Science, Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213, USA
{mrogati, yiming}@cs.cmu.edu

ABSTRACT

An under-explored question in cross-language information retrieval (CLIR) is to what degree the performance of CLIR methods depends on the availability of high-quality translation resources for particular domains. To address this issue, we evaluate several competitive CLIR methods - with different training corpora - on test documents in the medical domain. Our results show severe performance degradation when using a general-purpose training corpus or a commercial machine translation system (SYSTRAN), versus a domain-specific training corpus. A related unexplored question is whether we can improve CLIR performance by systematically analyzing training resources and optimally matching them to target collections. We start exploring this problem by suggesting a simple criterion for automatically matching training resources to target corpora. By using cosine similarity between training and target corpora as resource weights we obtained an average of 5.6% improvement over using all resources with no weights. The same metric yields 99.4% of the performance obtained when an oracle chooses the optimal resource every time.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *selection process;* H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *dictionaries*

General Terms

Algorithms, Measurement, Performance, Experimentation.

Keywords

Cross-language information retrieval, domain-specific translation

1. INTRODUCTION

Cross-language information retrieval (CLIR) has become increasingly important in information retrieval and related areas, including topic detection and tracking, question answering etc. Substantial progress has been made in both algorithm

development and evaluation methodology, and recent results in benchmark evaluations have shown that the performance of some CLIR systems has reached that of monolingual retrieval, as seen in TREC, NTCIR and CLEF [9][15][19]. For some researchers, these observations have lead to the optimistic conclusion that the CLIR problem is basically solved. More specifically, the problem is considered solved if high-quality training resources (parallel text, online dictionaries, multi-lingual thesauri, etc.) and/or high-performance machine translation (MT) systems are available, and meeting that condition is more of an engineering effort rather than research.

We argue that the above conclusion does not hold in general. CLIR systems' proven ability to rank news stories might not transfer readily to other genres such as medical journal articles – a point also raised by [16]. Indeed, the impressive CLIR performance was typically observed in the following settings:

1) test documents were general-domain news stories (i.e. those in TREC, CLEF and NTCIR), for which quality MT systems have been developed and tuned with person-decades of knowledge-engineering efforts; or

2) high-quality, general domain training resources were manually chosen by evaluation forum organizers or system developers. This is relatively easy when available resources are scarce but would be non-trivial with a large number of resources of varying quality. Moreover, some of these resources might be mismatched to the domain of the test collection.

We are concerned with this possible mismatch because, once we move away from news and into more technical domains, disambiguation becomes more problematic, in that the most common translation is no longer the one we usually desire. For example, the word *agent* should have different translation probabilities when the target corpus consists of newspaper stories (i.e. more likely to be a person or entity) vs. medical domain documents (more likely to be a chemical). Challenges for domain-specific CLIR, in particular the problem of distinguishing domain-specific meanings, have been noted in [12]. We argue that these variations can be captured by successfully matching training resources to target corpora.

Until recently, crosslingual resources were scarce and choosing which one to use was a quick decision made by the system engineer. However, today more and more data is available in electronic form; bilingual web pages are harvested as parallel corpora [20][14] as the quantity of non-English data on the web

increases; online dictionaries of various qualities and in various domains become available; previously translated documents are automatically aligned, and time-aligned comparable news are published every day. These resources are different in size, quality, vocabulary, genre, other domain characteristics, and purchasing cost. They provide the potential of significantly enhancing the performance of corpus-based statistical methods for CLIR and MT, if we have a systematic approach to the analysis of resources and an automatic solution for resource selection or weighting. To our knowledge, thorough investigation of this potential is not available in the literature of crosslingual information retrieval; our paper starts this investigation.

Several researchers addressed a somewhat related problem: choosing or weighting different translations, when more than one is available. A popular approach is to concatenate translations obtained from different sources. This does not take into account the target corpus and its domain, and it does not attempt to disambiguate query term translations. [4] combines the evidence for alternate translations by modifying Pirkola's structured query method to use translation probabilities. This approach does take into account target corpus characteristics, but it uses already unified translation probabilities that match the target corpus usage. In this context, favoring a translation that is common in the target corpus, but improbable given the training corpus, is deemed undesirable. We argue that technical, domain-specific terms exhibit exactly this behavior when a general-purpose training corpus is used. Consequently, correctly disambiguating these terms by choosing the translation that is more common in the target corpus is *not* undesirable. We are trying to accomplish this goal by incorporating domain-specific information into the translation probability. [3] addresses the issue by retaining the top two translations that occur most frequently in the target collection. We approach this problem by considering all translations, but weighting them differently depending on their translation probabilities as calculated from a linear combination of translation resources.

Domain-specific language modeling has been used in speech recognition [11][23], with encouraging results. [10] used CLIR followed by MT to find domain-specific articles in a resource-rich language, in order to use them for language modeling in a resource-poor language. This research is similar in spirit to ours, but has a very different focus.

Our investigation addresses the following questions:

- To what extent the performance of crosslingual retrieval methods would vary in a domain that is radically different from the news domain? We explore a technical domain by using journal article abstracts from medical literature, and several methods that were competitive in benchmark evaluations: the SYSTRAN commercial MT system, a weighted approach on IBM statistical MT translation probabilities, a corpus-based method using chi-square as similarity scores, and another corpus-based learning method using Pair-wise Mutual Information (PMI). To our knowledge, the latter has not been previously used or evaluated in CLIR .

- To what extent (if at all) can we improve the crosslingual retrieval performance by automated selection (and weighting) of the training resources,

compared to the natural (but optimistic) choice of using all available resources equally

- What criteria and algorithms can we use for optimizing resource selection automatically, based on the characteristics of both training resources and target collections?

We note that, while "resource selection" is a problem addressed in distributed IR, the similarity with the problem we are examining is only a matter of naming conventions. In distributed IR, the "resources" are (usually limited access) databases to be queried; here, our target collection *is* available and resources are defined as any aids in crossing the language barrier, including parallel corpora, dictionaries or MT systems. The methods and criteria presented in this paper are tailored to parallel corpora and dictionaries, since we are considering a specific technical domain.

The remainder of the paper is organized as follows: we present our training and testing data in Section 2, and our weighting criteria in Section 3. Section 4 discusses our CLIR approaches. Section 5 presents the results, Section 6 suggests future work, and Section 7 concludes.

2. EVALUATION DATA
2.1 Medical Domain Corpus: Springer
The Springer corpus is a product of the MUCHMORE project, an international effort concerned with cross-lingual retrieval in the medical domain. It consists of 9640 documents (titles plus abstracts of medical journal articles) in English and in German, with 25 queries in both languages, and relevance judgments made by native German speakers who are medical experts and are fluent in English. We split this parallel corpus into two subsets, and used the first subset (4,688 documents) for training, and the remaining subset (4,952 documents) as the test set in all our experiments. We applied an alignment algorithm to the training documents, and obtained a sentence-aligned parallel corpus with about 30K sentences in each language. The sentence-aligned version of the Springer training set was used in the experiments in this paper.

2.2 Training Corpora
In addition to Springer, we have used four other English-German parallel corpora for training:

- NEWS is a collection of 59K sentence-aligned news stories, downloaded from the web and covering the 1996-2000 period. It is available for download at http://www.isi.edu/~koehn/publications/de-news/

- WAC is a small parallel corpus obtained by mining the web, as described in [14]. It does not cover a particular domain.

- EUROPARL is a parallel corpus provided by [13]. Its documents are sentence-aligned European Parliament proceedings. This is a large collection that has been successfully used for CLEF, when the target corpora were collections of news stories [21].

- MEDTITLE, another product of the MUCHMORE project, is an English-German parallel corpus consisting

of 549K paired titles of medical journal articles. These titles were gathered from the PubMed online database (http://www.ncbi.nlm.nih.gov/PubMed/)

Table 1 presents a summary of the five training corpora characteristics. When showing the size in words, we include both languages.

Table 1. Characteristics of Parallel Training Corpora

Name	Approximate Size (sentences, words)	Domain
NEWS	59Kx2, 2M	news
WAC	60Kx2, 1.1M	mixed
EUROPARL	665Kx2, 35M	politics
SPRINGER	30Kx2, 0.9M	medical
MEDTITLE	550Kx2, 21M	medical

2.3 Data Preprocessing

We have eliminated all punctuation, stopwords and numbers for both languages. We used the Porter stemmer for English. For German, we first stemmed and we split the words into 5-grams to simulate German decompounding. The German stemmer and stopword list was provided by [22].

3. SELECTING TRAINING RESOURCES

We explore two domain-related criteria for selecting and weighting training resources. Other possible criteria are size, translation quality, and matching genre, but exploring them is beyond the scope of this paper, which focuses on showing why such selection is necessary.

3.1 Selection Criterion: Training Vocabulary Coverage

We begin by using vocabulary overlap between the training and testing corpora as a domain match approximation. Table 2 shows the vocabulary coverage with respect to the training collection. The training vocabulary coverage is calculated as follows:

$$Cov(train, test) = \left| V_{train} \bigcap V_{test} \right| / \left| V_{train} \right| \quad (1)$$

Table 2. Vocabulary coverage of training corpora with respect to the Springer test set

Name	DE Coverage (%)	EN Coverage (%)
NEWS	27.9	14.5
WAC	28.8	12.5
EUROPARL	6.6	1.8
SPRINGER	57.7	35.4
MEDTITLE	10.8	3.4

3.2 Selection Criterion: Cosine Similarity

The second simplest idea to approximate domain matching is using the cosine similarity between the testing and training corpus (TFIDF term weights). Note that these documents are very short (sentences), which means the document length is fairly constant and TF and DF tend to be close.

Table 3 shows the cosine similarity between the five training corpora and the test set.

Table 3. Cosine similarity of training corpora with respect to the Springer test set

Name	DE COS (%)	EN COS (%)
NEWS	44.08	33.90
WAC	54.67	33.84
EUROPARL	49.82	36.22
SPRINGER	99.29	90.89
MEDTITLE	55.20	72.94

3.3 Combining Translation Resources

Our proposed approach is to use vocabulary coverage or cosine similarity as the weight for each translating resource when combining them, instead of using it as a criterion for the selection threshold. Each parallel corpus produces a similarity matrix, using one of the methods outlined in section 4. A new similarity matrix is produced from their linear combination, using the vocabulary coverage or cosine similarity as the corresponding weights. In practice, it is only necessary to calculate this linear combination for dictionary entries present in the queries.

This approach has the advantage that it does not require relevance judgments and existent queries to learn the weights. We will examine the robustness of this approach in the Results section.

4. CLIR METHODS

We outline the corpus-based CLIR methods and a MT-based approach, with pointers to the literature where detailed descriptions can be found.

Let L1 be the source language and L2 be the target language in CLIR, all our corpus-based methods consist of the following steps:

1. Expanding a query in L1 using blind feedback

2. Translating the query, while preserving the weights from 1.

3. Expanding the query in L2 using blind feedback

4. Retrieving documents in L2

Here, blind feedback is the process of retrieving documents and adding the terms of the top-ranking documents to the query for expansion. We used simplified Rocchio positive feedback as

implemented by Lemur [18]. Our corpus-based methods differ only in the translation step, as described below.

4.1 Weighted Model 1 (WM1)

IBM's statistical machine translation Model-1 (or simply "Model 1") [1] uses a sentence-aligned training corpus to compute the term-term translation probabilities across two languages. The translation probability from term s (in L1) to term t (in L2) is defined as:

$$p(t \mid s) = \lambda_t^{-1} \sum_a P(S_s, a \mid S_t) \sum_{j=1}^{m} \delta(s, s_j) \delta(t, t_{a_j}) \quad (1)$$

where λ_t is a normalization factor, a is an alignment of cross-lingual term-term translation, S is a sentence in the L1 or L2 half of the parallel corpus, m is the number of tokens in the sentence in L1, and the second summation is the number of times s aligns with t in the corresponding alignment.

A matrix of translation probabilities is initialized and updated iteratively (we set the iteration number to 10 in our experiments). We use the resulting matrix to translate each query from L1 to L2: for each query word in the source language (L1), the entire vector of the corresponding target terms (in L2) is used in the translation, with the normalized probability as the weight of each target term. We named this method "Weighted Model 1" to distinguish it from using only the top target word in the translation of each source word.

Our approach is similar to IBM's and BBN's approaches to CLIR [6][7] except that the translation is not integrated in the retrieval model; only the query is translated. We found that this method performed well in CLIR benchmark evaluations [21].

4.2 Chi-square Statistic (CHI)

Chi-squared statistics are commonly used to measure the lack of independence between terms and categories in text classification; we are including it as a measure for term-term similarity between a source language term (s) and a target language term (t). CHI measures the dependence between s and t using four counts: A, B, C and D, where A is the number of passages (sentences or documents, depending on how the parallel training corpus is aligned) in which s and t co-occur, B is the number of passages s occurs without t, C is the number of passages t occurs without s, and D is the number of passages where none of them occur:

$$\chi^2(s,t) = \frac{(A+B+C+D) \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad (2)$$

This calculation results in a matrix of term-term associations, which we use for query translation in the same manner as the matrix of translation probabilities in WM1. The advantage of this calculation is its efficiency, compared to that of WM1. This makes it worth finding how effective CHI is in CLIR when compared to WM1.

4.3 Point-wise Mutual Information (PMI)

Point-wise mutual information is another common choice for measuring the empirical association between two variables (in our case, two terms across languages). The metric is defined as:

$$PMI(s,t) = P(s,t) \log \frac{P(s,t)}{P(s)P(t)} \quad (3)$$

The main difference between CHI and PMI is that PMI measures the positively correlated dependence while CHI counts both the positively and negatively correlated dependences. With respect to our task, translating a term from one language to another, PMI appears to be a more appropriate measure since we do not want to consider t as a translation of s if the joint probability of the two terms in human translations is too low. Comparative evaluation of PMI and CHI in CLIR was not reported before. In terms of computation, the two methods are equally efficient since the joint and marginal probabilities used in computing PMI can be easily derived from the counts of A, B, C and D defined in 4.2.

4.4 Weighted SYSTRAN (WSYS)

Although not a corpus-based method, we are including this approach in order to provide a comparison with a general-purpose machine translation system that is known to perform well in standard evaluation benchmarks such as CLEF [19]. We use SYSTRAN online to translate each query after the expansion using local feedback. In order to have a fair comparison, and not put SYSTRAN at a disadvantage, we preserve the term weights before the translation, and propagate the weight of each word to its translations. Post-translation query expansion is also included in the process and is identical to that of our corpus-based methods. Note that, unlike in the case of our corpus-based methods, morphological processing of a query has to be postponed until the query is translated.

5. RESULTS AND DISCUSSION

We conducted multiple sets of evaluations, all on the Springer test set. The results were evaluated using mean average precision (AvgP), a standard performance measure for IR evaluations, defined as the mean of the precision scores computed after each relevant document is retrieved (averaged over queries).

5.1 Empirical Settings

For the retrieval part of our system, we adapted Lemur [18] to allow the use of weighted queries. The retrieval model we used was simple TFIDF, the same used for our CLEF experiments in [21]. We used the publicly available software GIZA++ [17] as an implementation of IBM Model 1 [1]. Although more sophisticated translation models are also offered in GIZA++, we did not use them for this paper, for reasons of both efficiency and simplicity (e.g., word order is not our primary concern here).

Several parameters were tuned, naturally none on the Springer test queries. In the corpus-based approaches, the main parameters are those used in query expansion based on pseudo-relevance, i.e., the maximum number of documents and the maximum number of words to be used, and the relative weight of the expanded portion with respect to the initial query. Since the Springer training set is fairly small, setting aside a subset of the data for parameter tuning

was not desirable. We instead tuned the parameters on the CLEF collection[19]. Specifically, we chose 5 and 20 as the maximum numbers of documents and words, respectively. The relative weight of the expanded portion with respect to the initial query was set to 0.5.

In the following sections, DE-EN refers to retrieval where the query is in German and the documents in English, while EN-DE refers to retrieval in the opposite direction.

5.2 Can we use SYSTRAN directly in the Medical Domain?

To test whether CLIR systems that perform well in the news stories domain are robust enough to simply be used in a different domain, we have compared SYSTRAN (easiest, most convenient choice that worked extremely well in past evaluation forums) and two corpus-based methods trained on the Springer corpus. Figure 1 shows the results. As expected, although corpus-based methods performed close to the monolingual baseline (MLIR), SYSTRAN had difficulties with the specialized vocabulary, resulting in a drastic performance drop. Using the paired t-test, the difference between WSYS and CHI is statistically significant (p<0.005).
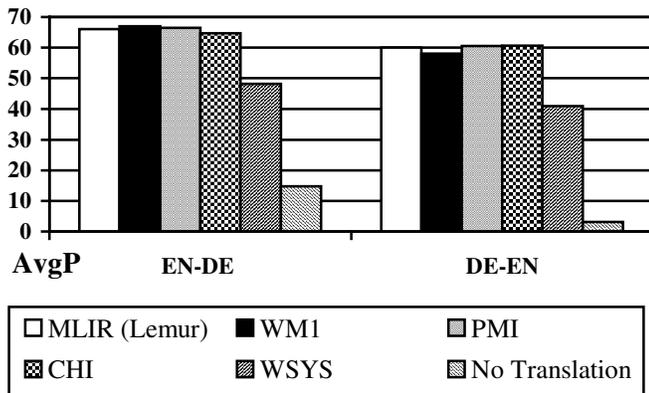

Figure 1. Retrieval Performance of corpus-based methods and Weighted SYSTRAN on the Springer test set

5.3 Is Resource Selection Necessary?

Given that we need to use corpus based approaches, can we simply choose a parallel corpus that performed very well on news stories, hoping it is robust across domains? Other natural approaches include choosing the largest corpus available, or using all corpora together. Figure 2 shows the effect of these strategies.
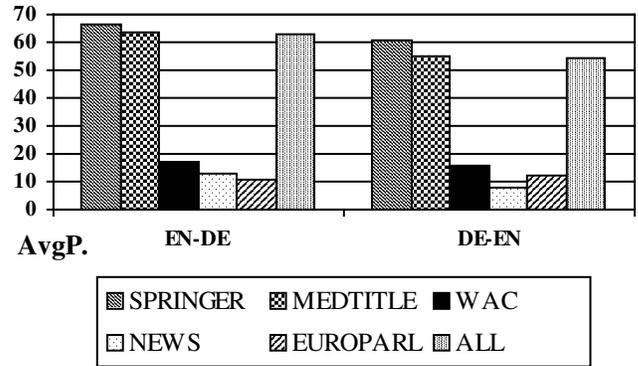

Figure 2. CLIR results on the Springer test set by using PMI with different training corpora

The experiments in Figure 2 and in the rest of the paper are performed using PMI, since it is faster than WM1 and performs equally well on Springer. We notice that choosing the largest collection (EUROPARL), using all resources available without weights (ALL), and even choosing a large collection in the medical domain (MEDTITLE) are all sub-optimal strategies. The difference between SPRINGER and ALL for DE-EN is statistically significant within a 95% confidence interval, and so are the differences between {SPRINGER, MEDTITLE, ALL} and {NEWS, WAC, EUROPARL}. Given these results, we believe that resource selection and weighting is necessary. We cannot simply use a well-performing, general-purpose parallel corpus for a technical domain. Thoroughly exploring weighting strategies is beyond the scope of this paper and would involve collection size, genre, and translation quality in addition to a measure of domain match. However, we propose cosine similarity as a simple weighting criterion and we examine this criterion as well as vocabulary coverage below.

5.4 Using Vocabulary Coverage or Cosine Similarity as Weights

Figure 4 and 5 show that both vocabulary coverage and cosine similarity between the training and target corpus are positively correlated with retrieval performance, as measured by the average precision over all queries. However, MEDTITLE is a significant outlier in this respect (when considering vocabulary coverage in Figure 4). We believe this to be because of the different spelling normalization applied to MEDTITLE when it was downloaded from the web. Many common 5-grams had a slightly different spelling. However, this disparity is alleviated in Figure 5, because the high TFIDF terms are medical terms where fewer spelling variations occurred.
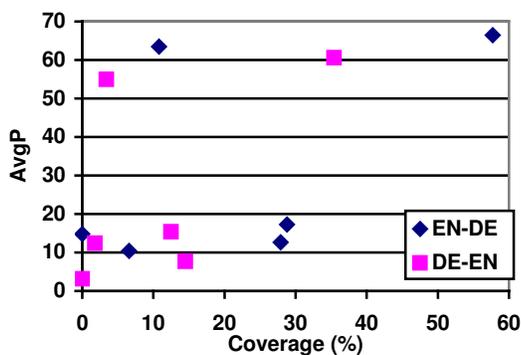
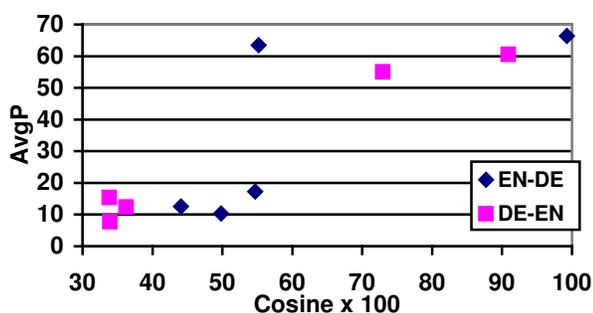**Figure 4. The CLIR performance (of PMI) vs. the training-set vocabulary coverage**

Figure 5. The CLIR performance (of PMI) vs. the cosine similarity

We wish to examine the robustness of using vocabulary coverage or cosine similarity between the target collection and the parallel corpus as linear combination weights. To that end, it is insufficient to experiment with weighting the five corpora mentioned above. For this approach to be robust, the combination should be consistent in not performing significantly worse than the best collection available – as given by an oracle - and in not performing worse than using all collections with equal weight. Naturally, the combination should also perform better than the expected performance of a randomly selected collection, but in our case this straw man baseline is not needed.

In Table 4 and its corresponding experiments, we simulate the availability of 5, 4 and 3 training collections from the five we have described above. We did not examine resource pairs, since the resource selection problem becomes trivial in this case. There are a total of 16 testing conditions: 1 way to choose all 5, 5 ways to choose 4, and 10 ways to choose 3 (order does not matter).

The first column enumerates which training collections we are allowed to select from and/or to weight. The respective collections (WAC, MEDTITLE, SPRINGER, EUROPARL, NEWS) are represented by their initials. All 16 combinations are shown for both DE-EN and EN-DE.

In this table, COV represents performance when training set coverage was used as the weight; COS represents performance when the cosine similarity was used as the weight; EQ represents using all available resources with equal weights; and *Best Single Collection* shows the performance of the single best one training corpus, if it were possible to know which one is the best in advance. Numbers in bold highlight the best performance of the four conditions. The star indicates statistical significance within the 95% confidence interval, using the paired t-test.

From this table, we see that cosine similarity is a better weighting measure than vocabulary coverage, which is to be expected. This simple method is robust when we simulate the availability of different collections; however, we believe more research is needed to explore different weighting criteria.

Summarizing across all 16 testing conditions, we observe that our strategy accounts for a 4-5% improvement over using all resources with no weights, for both retrieval directions. It is also very close to the "oracle" condition, which chooses the best collection in advance.

6. FUTURE WORK

We are currently exploring weighting strategies involving collection size, genre, and estimating translation quality in addition to a measure of domain match. Another question we are examining is the granularity level used when selecting resources, such as selection at the document or cluster level.

Similarity and overlap between resources themselves is also worth considering while exploring tradeoffs between redundancy and noise. We are also interested in how these approaches would apply to other domains.

7. CONCLUSIONS

We have examined the issue of selecting appropriate training resources for cross-lingual information retrieval. Our contributions include:

- We have shown that general-purpose MT systems and parallel corpora do not perform well when the target collection is in a domain with a highly technical vocabulary, such as medicine.

- We have introduced a new CLIR method (PMI), and evaluated the performance of several CLIR methods on a medical domain collection.

- We have evaluated the performance of several training parallel corpora on a medical domain collection, showing the need for properly matching the training and testing domains.

- We have proposed a simple and robust method for combining parallel corpora, based on domain match as approximated by cosine similarity.

Our results show that choosing the largest collection, using all resources available without weights, and even choosing a relatively large parallel collection in the medical domain are all sub-optimal strategies. Given these results, we believe that resource selection and weighting is necessary and can bring

significant performance improvements, especially within the domain-specific CLIR framework.

8. ACKNOWLEDGEMENTS

Table 4 . Results of CLIR runs using Various Resource Selection Strategies

Available Resources	DE-EN				EN-DE			
	COV	COS (%improvement over EQ)	EQ	Best Single Collection (Oracle)	COV	COS (%improvement over EQ)	EQ	Best Single Collection (Oracle)
WESNM	59.61	57.34(5.52%)	54.34	**60.56(S)**	66.36	**66.52 (5.23%)**	63.21	66.47(S)
ESNM	60.33	59.37(4.23%)	56.96	**60.56(S)**	66.51	**67.5 (2.61%)**	65.78	66.47(S)
WENM	40.24	**55.26(5.94%)**	52.16	55.1(M)	48.46	61.62(2%)	60.41	**63.5(M)**
WESM	59.07	57.7(5.61%)*	54.63	**60.56(S)**	66.92	**67.13(4.75%)**	64.08	66.47(S)
WESN	60.15	58.79(9.74%)*	53.57	**60.56(S)**	65.35	59.47(9.64%)	54.24	**66.47(S)**
WSNM	59.89	57.88(3.67%)	55.83	**60.56(S)**	66.71	**67.44(5.49%)**	63.93	66.47(S)
WNM	40.71	54.83(2.29%)	53.60	**55.1(M)**	52.49	62.26(-0.24%)	62.41	**63.5(M)**
WES	59.77	58.82(7.49%)	54.72	**60.56(S)**	65.74	64.91(19.76) *	54.20	**66.47(S)**
WEN	18.79	**19.01(0.21%)**	18.97	15.39(W)	22.22	**22.83(15.88)**	19.70	17.26(W)
WSM	59.27	57.25(3.11%)	55.52	**60.56(S)**	66.33	**66.86(2.76)**	65.06	66.47(S)
ESM	60.32	58.87(0.56%)	58.54	**60.56(S)**	66.98	66.44 (-1.24%)	**67.28**	66.47(S)
WEM	44.01	**55.42(2.74%)**	53.94	55.1(M)	56.07	61.78(0.29%)	61.60	**63.5(M)**
ESN	**60.88**	60.08(4.63%)	57.42	60.56(S)	**66.72**	66.56(11.02%) *	59.95	66.47(S)
SNM	60.31	59.57(1.72%)	58.56	**60.56(S)**	67.10	**67.19(0.97%)**	66.54	66.47(S)
ENM	48.39	**55.55(0.34%)**	55.36	55.1(M)	51.25	63.15(0.66%)	62.73	**63.5(M)**
WSN	59.86	59.68(7.33)	55.60	60.56(S)	65.08	65.02(10.74%) *	58.71	**66.47(S)**
Summary		**4.07%** average improvement over EQ		COS = **98.1%** Best		**5.65%** average improvement over EQ		COS = **99.4%** Best

9. REFERENCES

[1] Brown, P.F, Pietra, D., Pietra, D, Mercer, R.L. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19 (1993) 263-312

[2] Carbonell J. G, Yang, Y., Frederking, R. E., Brown, R., Geng, Y., Lee, D. Translingual Information Retrieval: A Comparative Evaluation. *In Proceedings of the IJCAI* (1) 1997: 708-715

[3] A. Chen, H. Jiang, and F. Gey. Combining Multiple Sources for Short Query Translation in Chinese-English Cross-language Information Retrieval. *In Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages, Sept. 30-Oct 1, 2000.*

[4] Darwish, K. and Oard, D. CLIR Experiments at Maryland for TREC-2002: Evidence Combination for Arabic-English Retrieval. *In TREC 2002 Proceedings*.

[5] Franz, M., McCarley, J. S, and Roukos, S. Ad hoc and multilingual information retrieval at IBM. *In The Seventh Text REtrieval Conference*, pages 157--168, November 1998. NIST Special Publication 500-242

[6] Franz, M. and McCarley, J.S. Arabic Information Retrieval at IBM. *In TREC 2002 proceedings*

[7] Fraser, A., Xu, J., Weischedel, R. 2002. TREC 2002 Cross-lingual Retrieval at BBN. *In TREC 2002 proceedings*

[8] Gey, F. and Jiang H. 1999. English-German cross-language retrieval for the GIRT collection – Exploiting a multilingual thesaurus. *In TREC-8 proceedings*.

[9] Kando, N. Overview of the Third NTCIR Workshop. *Working notes of the Third NTCIR Workshop Meeting. Part I:Overview*. Tokyo. Japan. October 2002. p.1-16

[10] Khudanpur, S., Kim, W., 2002. Using cross-language cues for story-specific language modeling. In *Proceedings of the International Conference on Spoken Language Processing*, p. 513-516

[11] Khudanpur, S. Kim, W., 1999. A maximum entropy language model to integrate n-grams and topic dependencies for conversational speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 553-556

[12] Kluck, M and Gey, F. The Domain-Specific Task of CLEF - Specific Evaluation Strategies in Cross-Language Information Retrieval. In C. Peters(Ed.), *Proceedings of the CLEF 2000 evaluation forum.*

[13] Koehn, P. Europarl: A Multilingual Corpus for Evaluation of Machine Translation. Draft, Unpublished.

[14] Nie, J. Y., Simard, M. and Foster, G.. Using parallel web pages for multi-lingual IR. In C. Peters(Ed.), *Proceedings of the CLEF 2000 evaluation forum.*

[15] Oard, D. W. and F. Gey, The TREC-2002 Arabic/English CLIR Track. *In TREC 2002 proceedings*

[16] Oard, D. When You Come to a Fork in the Road, Take It: Multiple Futures for CLIR Research. *Cross-Language Information Retrieval: A Research Roadmap.* Workshop at SIGIR-2002, Tampere Finland August 15, 2002

[17] Och, F. J. and Hermann N. Improved Statistical Alignment Models. In *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, (2000) pp. 440-447

[18] Ogilvie, P. and Callan, J. Experiments using the Lemur toolkit. In *Proceedings of the Tenth Text Retrieval Conference (TREC-10)*. (2001)

[19] Peters, C. Results of the CLEF 2003 Cross-Language System Evaluation Campaign. *Working Notes for the CLEF 2003 Workshop*, 21-22 August, Trondheim, Norway

[20] Resnik, P. Mining the Web for Bilingual Text. In *37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, College Park, Maryland, June 1999.

[21] Rogati, M and Yang, Y. Multilingual Information Retrieval using Open, Transparent Resources in CLEF 2003 . In C. Peters (Ed.), *Results of the CLEF2003 cross-language evaluation forum*

[22] Savoy, J. A stemming procedure and stopword list for general French corpora. *Journal of the American Society for Information Science,* 50(10) (1999) 944-952

[23] Seymore, K., Rosenfeld, R. 1997. Using story topics for language model adaptation. In *Proceedings of the European Conference on Speech Communication and Technology*