University of Chicago at CLEF2004: Cross-language Text and Spoken Document Retrieval

Gina-Anne Levow and Irina Matveeva

University of Chicago, Chicago, IL 60637, USA
levow,matveeva@cs.uchicago.edu,
WWW home page: http://people.cs.uchicago.edu/~levow

Abstract. The University of Chicago participated in the Cross-Language Evaluation Forum 2004 (CLEF2004) cross-language multilingual, bilingual, and spoken language tracks. Cross-language experiments focused on meeting the challenges of new languages with freely available resources. We found that modest effectiveness could be achieved with the additional application of pseudo-relevance feedback to overcome some gaps in impoverished lexical resources. Experiments with a new dimensionality reduction approach for re-ranking of retrieved results yielded no improvement, however. Finally, spoken document retrieval experiments aimed to meet the challenges of unknown story boundary conditions and noisy retrieval through query-based merger of fine-grained overlapping windows and pseudo-feedback query expansion to enhance retrieval.

1 Introduction

The University of Chicago participated in cross-language Multilingual, Bilingual, and Spoken Document Retrieval tasks. Cross-language experiments focused on meeting the challenges of new languages with freely available resources. We found that modest effectiveness could be achieved with the additional application of pseudo-relevance feedback to overcome some gaps in impoverished lexical resources. Experiments with a new dimensionality reduction approach for re-ranking of retrieved results yielded no improvement, however. Finally, spoken document retrieval experiments aimed to meet the challenges of unknown story boundary conditions and noisy retrieval through query-based merger of fine-grained overlapping windows and pseudo-relevance feedback query expansion to enhance retrieval.

2 Cross-language Multilingual and Bilingual Retrieval

The University of Chicago participated in the CLEF2004 cross-language multilingual and bilingual retrieval tasks. The group submitted four official English → English, French, Finnish, Russian multilingual runs, three using only the title-based topic specification and one using both the title and description components of the topic specification. The group also submitted one official English → Russian bilingual run in the title only condition. Additional unofficial contrastive runs discussed below highlight the effects of different processing.

2.1 Linguistic Resources

All processing employed only freely available linguistic resources. Two main classes of linguistic resources were utilized: bilingual term lists to bridge the gap between the information need as expressed in one language and the document concepts expressed in another and stemmers to perform simple morphological analysis to improve matching by reducing surface variation between information need and document concept forms. We downloaded bilingual term lists from http://www.freedict.com and Porter-style rule-based stemmers from http://snowball.tartarus.org. The overall size and coverage statistics for the bilingual term lists appear in Table 1. The English-French and English-Russian bilingual term lists provide an average of 1.5 translations for each English language term, while the English-Finnish lexicon averages approximately 1.2 translations. Although the French and Russian term lists are of comparable size, Finnish term list, in contrast, is relatively impoverished, being only one-tenth the size of the other term lists, providing translations for approximately 2500 English terms.

Lexicon	English Terms	Total Translations
English-French	21041	34949
English-Finnish	2546	3177
English-Russian	22722	31771

Table 1. Bilingual Term List Statistics

2.2 Document and Query Processing

We adopted a dictionary-based query translation architecture for all our runs to facilitate relatively rapid experimentation in a range of conditions. We applied comparable basic processing to all languages and conditions. Where specialized language specific processing was required, it is introduced in the detailed discussion below.

Document Processing Our goal in document processing was to reduce surface variation to enable matching with translated query forms or base queries in the case of English. All document languages undergo some morphological processing, although that of English is arguably simplest. Thus we applied the appropriate language-specific Snowball stemmer to each of the French, Finnish, and Russian document collections. Finally, all remaining accents were stripped. For English we relied on the INQUERY[1] retrieval system's built-in *kstem* stemmer.

For Russian and Finnish, some additional processing was required. For Russian, differences in coding formats required first conversion from the original document encoding to that correctly interpreted by the stemmer. Subsequently, to produce the 8-bit clean coding required by the retrieval engine, we produced an acceptable transliterated form. All Russian coding conversions employed the freely available *rucnv* (http://litwr.boom.ru) program.

Since Finnish is a highly agglutinative language, we also aimed to further reduce surface variation and identify suitable units for matching by decompounding. Specifically, we applied a greedy dictionary-based decompounder originally developed for previous experiments with German to split longer terms into word units attested by the translation resource.[2]

Query Processing Our query processing involves two phases: the first, from term extraction through translation, involves matching terms in the query with terms in the translation resource, while the second, following translation, involves matching with the target language documents and thus conforming to the earlier document processing.

First, based on left-to-right, greedy longest match, we identify multi-word units in the query that are translatable given the bilingual term list. Next we apply pre-translation pseudo-relevance feedback query expansion to enrich the short query with additional topically relevant and, we hope, translatable terms. For pre-translation expansion, we use the English document collection itself as a source of relevant documents and enriching terms. Starting with the original query formulation, with a default stopword list and stemming but no additional stop structure, we use the INQUERY API to identify the top ten presumed relevant documents from the collection and to identify terms more likely to appear in relevant documents than non-relevant documents. These terms are concatenated to the original query.

Next we perform dictionary-based term-for-term translation using the appropriate bilingual term list. We apply a stemming backoff procedure where we first attempt to match the surface form from the query with surface forms in the term list. Only if there is no match, do we back off to matching stemmed forms of query terms with stemmed forms of term list entries. We integrate evidence from all translation alternatives using structured query formulation as proposed by [3].

Now to support matching with the target language documents, we perform analogous processing on the translated queries to that performed on the documents. Specifically, we stem as described above, and perform language specific coding conversion for Russian and decompounding for Finnish. Finally, to further enrich the query and compensate for variation in choice of expression by document authors, we also perform post-translation pseudo-relevance query expansion. We apply a comparable mechanism to that for pre-translation expansion.

However, here we use the corresponding target language document collection as a source of both relevant documents and enriching terms.

2.3 Indexing and Retrieval

For baseline indexing and retrieval, we utilize the INQUERY information retrieval system version 3.1p1 licensed from the University of Massachusetts[1]. For each target document collection, we return the top 1000 ranked retrieved documents.

Locality Preserving Projection-based Re-scoring We applied a dimensionality reduction technique, the locality preserving projection (LPP) as described below, to perform a local re-scoring of the most highly ranked document in the ranked list.

LPP Algorithm In many cases, including text and images, the data can be assumed to be intrinsically low-dimensional although the traditional representation puts it in a very high-dimensional space. There has been a considerable amount of theoretical research and empirical investigation of representing data as points on the underlying manifold [4–6]. One hopes to obtain better similarity information by using the distance on the manifold instead of the distance in the ambient space.

The Locality Preserving Projection (LPP) algorithm [7] computes a linear projection of the data that preserves the intrinsic geometric structure of the manifold. The input is n data points as vectors in R^N $X = (x_1, ..., x_n)$ that belong to a k-dimensional manifold M embedded in R^N. The goal to find a lower-dimensional representation for these points $y_1, ..., y_n \in R^k$, where $k < N$. First, a neighborhood graph of the data $G = (V, E)$ is constructed. W is the adjacency matrix of the graph. The entry W_{ij} is non-zero if the data points i and j are connected by an edge $e \in E$. The entries of W contain the information about the local similarities between the data points. The next step is to compute the diagonal matrix C of node's degrees, where $C_{ii} = \sum_{j=1}^{n} W_{ij}$, and the Laplacian of the graph $L = C - W$.

LPP finds a lower-dimensional representation $y_1, ..., y_n \in R^k$, where $k < N$ so as to minimize under certain constraints

$$\sum_{ij} ||y_i - y_j||^2 W_{ij}$$

where W_{ij} is the penalty on the distance between the points y_i and y_j. W_{ij} is large if the original points x_i and x_j corresponding to y_i and y_j are close. Thus, if data points are similar to each other in the input space, there will be a penalty for mapping them far apart and they will remain close to each other in the new representation.

It can be shown [7], that the solution is given by the generalized eigenvectors of the following generalized eigenvalue problem $XLX^T a = \lambda XCX^T a$.

A constraint is necessary to prevent the algorithm from collapsing all input vectors to just one point. Here we used the constraint $a^T a = 1$ and thus we had to solve the eigenvalue problem $XLX^T a = \lambda a$ to find the solution.

With any constraint, k (generalized) eigenvectors corresponding to the k smallest (generalized) eigenvalues form the projection matrix A_k. The new representation of the data is computed as $Y = A_k^T X$.

LPP re-ranking of the candidates list We made the assumption that the baseline system performed well, specifically that the top thousand documents in the ranked list that this system returned contained the relevant documents. Thus, we could apply LPP locally, only to these documents, avoiding the computational intractability of this technique for larger document and term spaces which did not permit us to apply the technique to the collection as a whole.

Preprocessing We used the following preprocessing steps. All documents from the Russian[1] part of the collection were used to compute the vocabulary as well as the term and document frequencies for the vocabulary terms. After that the top documents and queries were indexed and weighted using $tfidf$. We use the Rainbow document classification package [8] to perform the indexing.

LPP projection Using these top documents from the ranked list returned by the baseline system, we computed the LPP model:

- Using the Euclidean distance compute the nearest neighbor graph of the data
- Compute the graph Laplacian
- Compute the LPP projection vectors
- Using the LPP projection vectors, fold in the documents and the queries to obtain their low dimensional representation

The inner product between the new document and query vectors was used as the measure of their similarity. Using this similarity score, a new ranked list was computed.

LPP perspective We had the following motivation for using the LPP re-ranking. LPP is a dimensionality reduction algorithm and performs a certain kind of denoising. In the LPP space documents that are similar to each other in the original representation remain close. Thus, if some of the top documents in the ranked list returned by the baseline system were actually relevant to the query, LPP would map other documents that are placed at lower ranks close to the top ones. This can increase the rank of the other relevant documents.

[1] Due to time limitations, the LPP re-scoring was applied only to the Russian bilingual and Russian portion of the multilingual retrieval task.

Multilingual Merging Finally, since we perform query translation into multiple document languages for the retrieval in the multilingual task, it is necessary to merge the ranked lists from the individual per-language retrieval runs to produce a single ranked list. Based on a previous side experiment, we determined that there was a clear relation between number of untranslated terms in the final query formulation and the retrieval effectiveness of the query. Previous experience had indicated that fully enriched CLIR techniques could achieve retrieval effectiveness comparable to or even better than monolingual retrieval effectiveness due to implicit and explicit enrichment processes.

We assumed a rank-based, round robin merge strategy across the per-language runs, up to a total of 1000 documents in the final ranked list. Based on the potential high effectiveness of CLIR where translation was highly successful, we assumed a uniform merge strategy when full or almost full translation was achieved. On a per-query basis, we reduced the contribution of each per-language ranked list based on observed decreases in translation success. Based on the side experiments, we identified thresholds for full, partial, and poor translation success, as indicated by the residual presence of untranslated terms in the final query formulation. Each reduction in translation success level resulted in a reduction of one-third in the contribution of that language's ranked list to the final ranked list.

Merging was not necessary for the monolingual or bilingual runs.

2.4 Results and Discussion

We present the results for the merged multilingual runs. We also present contrastive bilingual results for specialized processing that was applied only to one document language or that had different effects across languages that might not yield significant effects at the merged multilingual level. We apply the Wilcoxon signed ranks test to assess statistical significance of differences between two sets of retrieval results.

Multilingual Runs We find that, relative to the baseline runs, decompounding for Finnish appears to enhance retrieval and the LPP re-ranking in Russian appears to decrease effectiveness (Table 2). These contrasts do not reach significance. Since these modifications affect only two of the target languages, it is not surprising that the changes do not lead to significant overall changes in effectiveness.

Query	Baseline	+Decompounding	+LPP Re-scoring
Title	0.1464	0.1545	0.1307

Table 2. Multilingual Runs

Bilingual Contrasts: Finnish Decompounding We find that relative to baseline effectiveness, changes due to Finnish decompounding did not reach significance.

Query	Baseline	+Decompounding
Title	0.1979	0.2207
Title+Description	0.2383	0.2308

Table 3. Effects of Finnish Decompounding

Bilingual Contrasts: LPP-based Re-scoring We find that relative to baseline effectiveness, LPP-based re-scoring fares significantly more poorly (Table 4). One possible contribution to LPP's failure to improve over baseline is the relatively small number of on-topic documents in the Russian collection, resulting in large effects for changes in a few document positions. Another possible explanation for LPP's failure to improve the retrieval performance is that the LPP projection was computed using the similarity between the documents themselves, not their similarity to the query. However, documents that are relevant to the same query are not necessarily similar to each other. It has even been observed that every query defines a new similarity notion between the documents. In the future we will consider applying a pseudo-relevance approach in which we explicitly presume the highest ranked documents to be relevant and adapt the connectivity graph as appropriate.

Query	Baseline	LPP
Title	0.1199	0.0029
Title+Description	0.1611	0.0021

Table 4. Effect of LPP Re-scoring

Bilingual Contrasts: Pre- and Post-translation Expansion We find an apparent trend to increases in effectiveness for pseudo-relevance feedback query expansion relative to retrieval without expansion(Table 5). However, we find that only for the Finnish case do these differences reach significance ($p < 0.01$). In particular, for Finnish, pre-translation expansion yields significant improvements over both no expansion and post-translation expansion alone. This large contrast can be best understood in the context of the highly impoverished - $\approx$2500 headword - bilingual term list available for Finnish. For comparison, the French and Russian term lists have almost ten times as many headwords. Thus pre-translation expansion plays a key role in enabling translation and matching of query concepts. This behavior is consistent with [9]'s prior findings on artificially impoverished translation resources.

Query	No Expansion	Post-expansion	Pre- and Post-expansion
FR Title	0.1300	0.1710	0.1656
FR Title+Description	0.1538	0.1843	0.1866
FI Title	0.1427	0.1505	0.2279
FI Title+Description	0.1610	0.1616	0.2308
RU Title	0.1051	0.0963	0.1199
RU Title+Description	0.1270	0.1201	0.1611

Table 5. Effects of query expansion

3 Spoken Document Retrieval

The University of Chicago also participated in the CLEF2004 cross-language spoken document retrieval task. Runs were submitted in both the baseline English monolingual task and the French-English cross-language task, using only the resources provided by CLEF.

3.1 Query Processing

Query processing aimed to enhance retrieval of the potentially errorful ASR transcriptions through pseudo-relevance feedback expansion. The baseline conditions required the use of only the CLEF provided resources. This restriction limited our source of relevance feedback to the ASR transcriptions, segmented as described below. For both the monolingual English and the English translations of the original French queries, we performed the same enrichment process. We employed the INQUERY API to identify enriching terms based on the top 10 ranked retrieved segments and integrated these terms with the original query forms. Our hope was that this enrichment process would capture both additional on-topic terminology as well as ASR-specific transcriptions.

For the French-English cross-language condition, we performed dictionary-based term-by-term translation, as described in [2]. We employed a freely available bilingual term list (www.freedict.com). After identifying translatable multi-word units based on greedy longest match in the term list, we used a stemming backoff translation approach with statistically derived stemming rules[10], matching surface forms first and backing off to stemmed forms if no surface match was found. All translation alternatives were integrated through structured query formulation[3].

3.2 Spoken Document Processing

This year the SDR track focused on the processing of news broadcasts with unknown story boundaries. This formulation required that sites perform some automatic segmentation of the full broadcasts into smaller units suitable for retrieval. Using an approach inspired by [11], we performed story segmentation as follows. First we created 30 second segments based on the word recognition time stamps using a 10 second step to create overlapping segment windows. These units were then indexed using the INQUERY retrieval system version 3.1p1 with both stemming and standard stopword removal.

3.3 Retrieval Segment Construction

To produce suitable retrieval segments, we merged the fine-grained segments returned by the base retrieval process on a per-query basis. For each query, we retrieved 5000 fine-grained segment windows. We then stepped through the ranked retrieval list merging overlapping segments, assigning the rank of the higher ranked segment to the newly merged segment. We cycled through the ranked list until convergence. The top ranked 1000 documents formed the final ranked retrieval results submitted for evaluation.

3.4 Results and Discussion

In Table 6, we present the results for both the monolingual baseline and the cross-language English → French spoken document retrieval runs in the unknown story boundary condition. There is a substantial drop-off in retrieval effectiveness for the cross-language relative to the monolingual runs. Post-hoc examination of the translated queries strongly suggests the need for addition stopword and stop structure removal for the French topics. There is also an apparent, but not significant, 10% relative improvement for the expanded French query over the unexpanded case. The effectiveness of the monolingual runs suggests the potential of spoken document retrieval in the unknown story boundary condition, even with a simple window merging approach to segmentation.

Query	Monolingual	French No-Exp	French Expanded
Description	0.2820	0.0885	0.0965

Table 6. Spoken Document Retrieval

4 Conclusion

In the CLEF2004 multilingual and bilingual experiments, we demonstrated the flexibility of a dictionary-based query translation architecture by extension to two new languages, Finnish and Russian, with freely available translation and stemming resources. We further found significant utility in pre-translation query expansion for a language with only a rudimentary translation resource, enabling translation of key concepts. Experiments with a locality preserving dimensionality reduction technique suggest future work in which the likely relevance of the highest ranked documents is used explicitly for result re-scoring. Finally the spoken document retrieval results suggest that even a simple window-based approach to segmentation can yield modest retrieval effectiveness. However, future research will explore augmenting the window-based segmentation approach with a richer topical, possibly query-independent, segmentation.

References

1. Callan, J.P., Croft, W.B., Harding, S.M.: The INQUERY retrieval system. In: Proceedings of the Third International Conference on Database and Expert Systems Applications, Springer-Verlag (1992) 78–83
2. Levow, G.A., Oard, D.W., Resnik, P.: Dictionary-based techniques for cross-language information retrieval. Information Processing and Management (to appear)
3. Pirkola, A.: The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (1998) 55–63
4. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality +reduction and data representation. Neural Computation **15** (2003) 1373–1396
5. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science **290** (2000) 2323–2326
6. Tenenbaum, J., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290** (2000) 2319–2323
7. He, X., Niyogi, P.: Locality preserving projections. In: Proceeding of NIPS 2003. (2003)
8. McCallum, A.K.: Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow (1996)
9. McNamee, P., Mayfield, J.: Comparing cross-language query expansion techniques by degrading translation resources. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (2002) 159–166
10. Oard, D.W., Levow, G.A., Cabezas, C.: CLEF experiments at the University of Maryland: Statistical stemming and backoff translation strategies. In: Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 2000, Revised Papers. Volume LNCS 2069 of Lecture Notes in Computer Science., Heidelberg, Springer-Verlag (2001) 176–187
11. Abberley, D., Renals, S., Cook, G., Robinson, T.: Retrieval of broadcast news documents with the thisl system. In Voorhees, E., Harman, D., eds.: Proceedings of the Seventh Text REtrieval Conference (TREC-7). (1999) 181–190 NIST Special Publication 500-242.