Combining Categorization-based and Corpus-based Approaches for CLIR

Yiming Yang, Monica Rogati and Bryan Kisiel

Carnegie Mellon University, Pittsburgh, Pennsylvania

{yiming, mrogati, bkisiel}@cs.cmu.edu

Abstract

Whether or not we can use existing concept taxonomies to help cross-lingual information retrieval (CLIR) is an open question. This paper investigates an interlingual approach that uses the MeSH categories in the medical domain to index bilingual documents and queries and to measure their relevance based on a category-level matching. We conducted bilingual retrieval experiments on a new corpus (Springer) of medical documents and queries, in the languages of English and German. We also evaluated several high-performing corpus-based learning methods and a machine translation (MT) based approach using SYSTRAN, a commercial system with strong results on CLIR benchmarks. Our results on Springer show that the categorization-based approach significantly outperformed the MT-based approach, but underperformed the corpus-based methods due to the loss of detailed information from the category-level indexing. Combining the output of categorization-based retrieval and corpus-based retrieval yielded a significant performance improvement over using either alone.

Introduction

Crosslingual information retrieval (CLIR), the problem of retrieving documents in one language using queries in another, has become an important area for research. Both corpus-based learning and machine translation techniques have been studied for crossing the language barriers in this task. The IBM Model-1 for statistical MT [1][4], for example, has been widely used to extract bilingual dictionaries from training corpora of parallel text. The SYSTRAN commercial machine translation system has also been a common choice for translating documents or queries. In recent years, substantial progress has been made in CLIR, as is evident in the benchmark evaluations conducted by TREC, CLEF and NTCIR, where the best CLIR systems often had results equally as good as — and sometimes even better than — the best performance in monolingual retrieval (MLIR)[12][14][20]

In contrast to the intensive publications on corpus-based and MT-based CLIR approaches [12][14][20], studies on categorization-based interlingual approaches are relatively rare – the closest is [8] , which uses EuroWordNet as an interlingua. By *categorization-based interlingua*, we mean a universal language that uses predefined categories or concepts as the semantic units to represent queries and documents in any language, as well as an algorithm to perform the mapping from the query or document languages to the interlingua. Other approaches to using a multilingual taxonomy as interlingua use words, not concepts, as the indexing unit, and the taxonomy as a dictionary [3][7][10].

Human-defined taxonomies of categories exist in many domains, and have been widely used in databases to index multilingual documents and to support user browsing. In the medical domain, for example, a large number (over 20,000) of Medical Subject Headings (MeSH) have been defined by professionals, organized into a hierarchy and treated as the semantic units in MEDLINE indexing of multilingual documents. It is natural to wonder whether we can build a CLIR system that uses MeSH as the interlingua that provides the mapping from multilingual queries and documents to the taxonomy of categories, and that matches queries and documents based on the bag-of-categories representations of these items. We wonder how well (or how poorly) such a system would perform compared to corpus-based learning and MT-based approaches. Answering this question requires a thorough investigation, which is the theme of this paper.

Our study consists of the following components:

- Proposing a categorization-based interlingual approach and examining it using MeSH categories and the Springer corpus, a new test bed of English-German documents and queries in the medical domain
- Generating CLIR performance baselines of corpus-based methods on Springer for comparison, including IBM Model-1, Pair-wise Mutual Information (PMI), Chi-square (CHI) statistic, and pseudo-relevance feedback (PRF), and also MT-based query translation using SYSTRAN
- Generating MLIR performance baselines on Springer for comparison by running a (word-based) high-performing retrieval system (Lemur)[19] and the categorization-based retrieval system with the same indexing language (MeSH) in monolingual settings
- Developing and evaluating a new approach to multilingual text classification that allows the use of categorized documents in one language to train a classifier for another language when labeled training documents are available for the first language but not

the second language, and when a training set of parallel documents (unlabeled) is available

- Optimizing CLIR performance by combining categorization-based and corpus-based approaches

Related Work

Eichmann et al.[3] studied the use of pre-existing multilingual translations (English, Spanish and French) of the MeSH categories to cross the language barriers in crosslingual retrieval. The query is augmented with the MESH terms associated with previously retrieved documents. Then, the English equivalent of these MESH terms would be added to the query, which is "refined" using several rules, and eventually treated as a bag of terms (words or phrases) in English. The average precision of this approach on the OHSUMED corpus was between 61%-75% of that of monolingual retrieval. An issue is the overlap assumption between the controlled MESH vocabulary and the terms used in the actual English documents -- those vocabularies only partially overlap. Gey and Jiang [7] reported a similar approach on the GIRT corpus in the field of social science, and Hersh and Donohoe [10] utilize an analogous algorithm, with the practical application of using it to retrieve previously classified medical documents in response to a query in a different language.

While the above work suggests useful ways of exploiting existing taxonomies for crosslingual retrieval, those approaches are fundamentally different from the categorization-based interlingual approach we are investigating. Neither of the approaches is truly interlingual in the sense that categories were not used as the semantic units in the indexing language; instead, individual terms (including single words, noun phrases, etc.) in the document language were used as the units for indexing documents and for matching documents and queries. In other words, existing thesauri of categories were used as a means for query translation from one language to another, not as an interlingua language for concept-level representations. However, concept-level indexing has been explored by [8], using the language independent EuroWordNet synsets/concepts as indexing units. A classifier is not needed here, since the mapping between the words and concepts is given by EuroWordNet, although disambiguation is a problem since the mapping is not unique. Known-item retrieval evaluation on 171 documents and their summaries used as queries showed that synset indexing performed about the same as word-based retrieval, unless manual disambiguation was provided. Although our evaluation is limited to the medical domain, our test collection is closer to the standard TREC-style evaluation, in number of documents, query generation method and human relevance feedback.

These studies, therefore, cannot satisfactorily answer the questions of whether using concepts designed by humans as indexing units can help cross-lingual retrieval, and to what degree text categorization can be effectively used in such a process. To our knowledge, these questions have not been answered in CLIR research; exploring the answers is the main contribution for which we aim in this paper.

Another major distinction of the above approaches from the one we present is that they rely on the availability of multilingual translations of categories. According to Eichmann et al., only 13.7% of the MeSH concepts had Spanish translations in UMLS at the time [3]. This means that the majority of MeSH concepts were not well covered by the reduced controlled vocabulary in Spanish. This is also true of EuroWordNet. A categorization-based approach that uses statistical learning (as ours does), on the other hand, does not require any textual annotation of categories. As long as some categorized documents are available for each language, a mapping from free vocabularies to the interlingua can be learned automatically by the system. Additionally, we also explore a novel approach to multilingual classification for CLIR, i.e. using a parallel corpus to generate a "pseudo-training set" for a language when human-labeled documents are not available for training. No literature that we are aware of has examined this approach.

Categorization-based Approach

Our approach consists of two steps: 1) classifying queries and documents into categories, and 2) retrieving documents for queries in the category space. For the first step, we train statistical classifiers for the query language and the document language. These classifiers are responsible for mapping queries and documents from vectors of weighted terms to vectors of weighted categories (in our case, MeSH concepts). Once both queries and documents are mapped onto vectors in the category space, their original representations become irrelevant, and the remaining part of the process is the same as retrieval in conventional vector space. We use a publicly available retrieval system (Lemur , see Section 6) for this part of the approach.

The key question is how to automatically learn a mapping from each language (source or target) to the category space. We chose to use k-nearest neighbor (kNN) classification, a well understood and high performing algorithm in text categorization evaluations [15]. Depending on the availability of the training data, we explore two alternatives for the training process: straightforward training and cross-language transitive labeling.

Straightforward Training

If we have a training set of categorized documents in both the source language and the target language, we apply the standard kNN procedure in each language. Considering an input query or document as a vector of weighted words, the system retrieves its k nearest neighbors (documents in the same language) from the training set, and uses the nearness of each neighbor as the weight of the categories of that neighbor. By collecting all the categories of the k nearest neighbors and summing their weights by category, we obtain a vector of weighted categories. More specifically, for term weighting we use a standard TF-IDF scheme (“ltc” in the

SMART nomenclature), and for nearness we use the standard cosine similarity. As a minor refinement, we applied local (same-language) feedback to each query or document before the classification process.

Cross-language Transitive Labeling

If we have a categorized training set only in one language, and if a parallel corpus is available, we propose a transitive labeling process:

- Use the categorized documents to build a kNN classifier for the first language (L1)
- Use the kNN classifier to assign categories to the documents in the parallel corpus based on the documents in the L1 half of the corpus
- Propagate the categories through the document pairs in the parallel corpus from L1 to the second language (L2)
- Use the system-classified documents in L2 as the training set to build a kNN for the second language

An example would explain why such a transitive classifier is desirable. For English, we have the OSHUMED collection (Section 5.2) as a labeled training set which is large in volume, with high-quality category assignments by trained professionals, and containing documents with a reasonable length of text (title plus abstract for each article). However, we could not find a similar training corpus in German. This situation motivated us to try the transitive labeling process as an alternative, taking advantage of availability of a categorized monolingual corpus and an un-categorized parallel corpus. Although we use kNN to test this approach in this paper, it can be replaced with any other classifier.

Corpus-based and MT-based Approaches

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

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

1. Expand query in L1 using local feedback
2. Translate the query
3. Expand query in L2 using local feedback; retrieve

Here local feedback is the process of retrieving documents and adding the terms of the top-ranking documents to the query for expansion; those documents are weighted using their cosine similarity scores computed against the query. Our corpus-based methods differ only in the translation step. For space reasons, we cannot include their description here, but they can be found in [22] for Weighted Model 1 (WM1), Chi-Square (CHI), Pointwise Mutual Information (PMI), Weighted Systran (WSYS) and in [26] for Cross-Lingual Pseudo-Relevance Feedback (CL-PRF).

Data Sets

Springer

The Springer corpus 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 fluent in English. We split this parallel corpus into two subsets, and used 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 silence-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. Category labels are not available in this collection, so it cannot be used for training a categorization-based system. The number of relevant documents per query is 18.8 on average in both the training and test sets.

OHSUMED

The OHSUMED corpus is a monolingual collection of documents (titles plus abstracts of medical journal articles), and a benchmark used in text categorization evaluations [14][11]. It consists of 233K medical abstracts with 14K categories in MeSH, and the number of categories per document is 13 on average. We use this corpus to train a kNN classifier for English directly, and we use it in combination with the Springer training set (parallel text, but not categorized) to obtain a kNN classifier for German through a cross-language transitive labeling process (Section 3.2). The documents in OHSUMED are from a five-year period, 1987 to 1991. In our experiments in this paper, we only used the 1987 portion of the data (36,890 documents, 10,889 categories).

MedTitle

MedTitle is an English-German parallel corpus consisting of 549K paired titles of medical journal articles. In our experiments for this paper, the parallel-text part of MedTitle was used as the training data in the corpus-based approaches, and titles plus categories were used as the training data in the categorization-based approach. That is, for the corpus-based approaches, the German titles were used to learn a mapping from German to MeSH, and the English titles were used to learn a mapping from English to MeSH.

Evaluation

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. It is defined as the mean of the precision scores computed after each relevant document is retrieved.

Empirical Settings

For the retrieval part of our system, we adapted Lemur [19]in the way that allows the use of weighted categories for document indexing and weighted documents for query

expansion. We also used the publicly available software GIZA++[18] 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, none on the test set. 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 [20]. Specifically, we chose 5 and 20 as the maximum numbers of documents and words. The relative weight of the expanded portion with respect to the initial query was set to 0.5.

The main parameter in the kNN classifiers is the value of *k*. According to previously reported experiments for kNN on OHSUMED, the suitable range for the value of k was around 50 to 200. We then set k=50 based on such knowledge; we did not tune this parameter. The relative weight between categorization- and corpus-based retrieval is a parameter when combining the two. We tuned this parameter on the training set.

Main Results

Figures 1 and 2 show the results of the corpus-based, the categorization-based and the MT-based approaches (Sections 3 and 4) on the Springer test set. Since the collection is bilingual, retrieval performance in both directions was evaluated: EN-DE means using English queries to retrieve German documents, and DE-EN means using German queries to retrieve English documents.

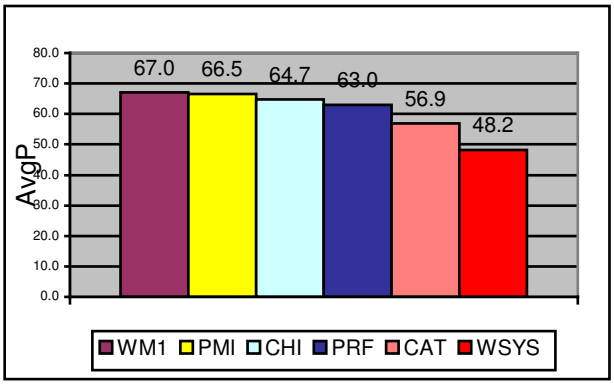

Figure 1. CLIR performance in EN-DE retrieval

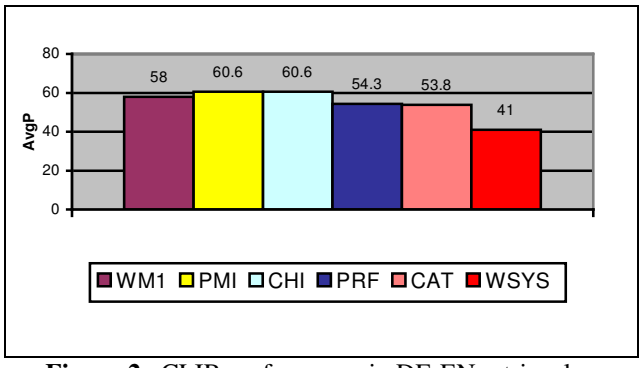

Figure 2. CLIR performance in DE-EN retrieval on Springer

All the corpus-based methods (WM1, PMI and CHI) were trained on the Springer training set, and the categorization-based method (CAT) was trained on the MedTitle corpus. The MT-based method (WSYS) did not use any training set but SYSTRAN. We are pleased to see the competitive performance of PMI and CHI compared to WM1, which has performed strongly in CLIR benchmark evaluations. CAT had the performance similar to CL-PRF in one direction (DE-EN) but worse than those of the corpus-based methods in other cases; however, it was significantly better than the MT-based approach in both directions. It is the first time for a categorization-based interlingual approach to be empirically evaluated in comparison with corpus-based and MT-based approaches in CLIR. The results are indeed encouraging (see the discussion in Section 6.6). The weak performance of WSYS is expected due to the technical domain.

Training Issues in CAT

Notice that the CAT results we used in the above comparison are those from using MedTitle corpus to train the classifiers (in English and German).

We also examined CAT with cross-language transitive labeling. That is, we used OHSUMED to build a kNN classifier in English, and the Springer training set for the transitive labeling and building a kNN classifier for German. Figure 3 compares the results of CAT under these two training conditions: direct (using MedTitle) and transitive (using OHSUMED plus the Springer training set). Notice that neither condition is "ideal": MedTitle consists of short text (titles) only, while OHSUMED does not have the German half. The experimental results show that MedTitle is a better choice, and that it is better even for just the English part of the task (not affected by the transitive labeling process. Notice that the transitive version of CAT still outperformed the MT-based approach using SYSTRAN.

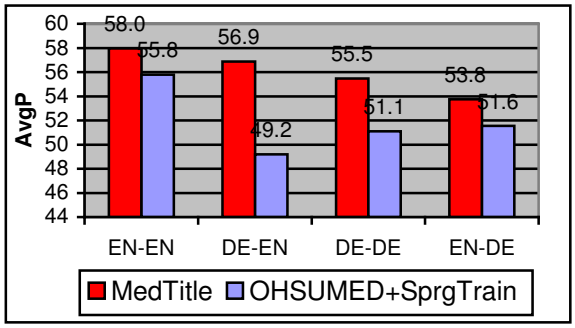

Figure 3. CAT with different training conditions

Effects of Combining Categorization-based and Corpus-based Approaches

We wonder whether we can improve CLIR performance by combining corpus-based and categorization-based approaches. Figure 4 shows the experimental results. We chose to use PMI in the combination experiments for its performance competitive with WM1 and its more efficient computation. We used the MedTitle corpus as the training set for both PMI and CAT. The performance improved in both directions when combining PMI and CAT. These improvements are statistically significant (p-values < 0.01), according to results of significance tests for proportions [27].

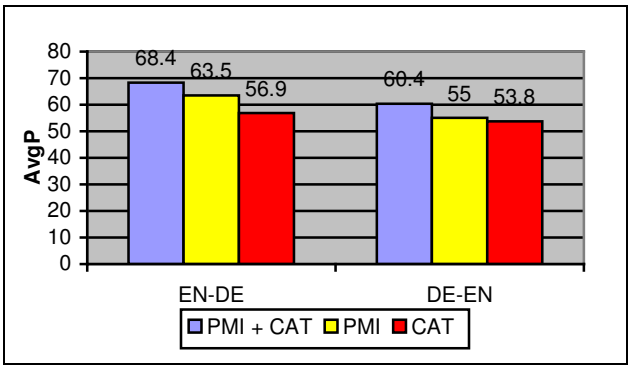

Figure 4. Effect of combining CAT and PMI in CLIR (both trained on MedTitle)

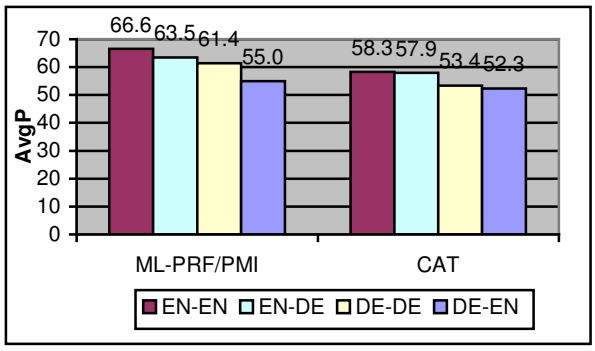

Figure 5. Performance comparison: MLIR vs. CLIR, and word vs. category-based representation

Monolingual Baselines

In order to obtain monolingual retrieval baselines for comparison, we ran the Lemur retrieval system with default tf.idf settings on the English and German parts of the Springer test set, respectively.

The CLIR performance of PMI was 95% (63.5/66.6) of that obtained by ML-PRF on English queries, and 90% (55.0/61.4) on German queries. The CLIR performance of CAT is 99% (57.9/58.3) of the MLIR performance of the same method on English queries, and 98% (52.3/53.4) on German queries. Figure 5 shows these comparisons.

CLIR performance of both PMI and CAT are highly effective for crossing the language barriers in CLIR. The performance difference between PMI and CAT comes mainly from the choice of data representation: CAT performed a mapping from word-based to a category-based vector representation; information in the original documents or queries may be lost.

Evaluation Methodology Issues

A potential limitation of the evaluations we presented in this paper is that we only focused on a standard retrieval task. What we are not measuring is whether a user can be benefited from being presented with the category-based representation of documents and being able to follow those categories in browsing. Given that category taxonomies were originally designed for organizing a large information space and supporting easy browsing by users, the evaluation methodology we followed in this paper, although "standard", may not be sufficient for assessing the full benefit of categorization-based retrieval support. User-oriented evaluations are clearly important for making a better assessment in this aspect. The slightly lower AvgP scores of CAT compared to those of the best corpus-based methods should not be interpreted as negative evidence for the benefit of using categorization-based retrieval. It is encouraging to see the close performance of categorization-based retrieval compared to term-matching based retrieval, in an evaluation framework that is particularly suitable to the latter.

Other experiments on the Springer corpus have been published in [25]. We found those not directly comparable with ours because the data used for training and the data used for evaluations were not well separated in those experiments. Also, the reported performance scores were too low (with AvgP scores up to .35) to serve as challenging baselines.

Concluding Remarks

The reported work can be summarized as follows:

- We conducted the first comprehensive empirical validation of a categorization-based interlingual approach to CLIR, examining it with MeSH (as the interlingua) in the medical literature domain and on the Springer bilingual corpus.
- We provided monolingual performance baselines on Springer as well, including results from the Lemur

system and our categorization-based system with MeSH-based indexing for monolingual retrieval.

- We developed and evaluated a new variant of our categorization-based method, the cross-language transitive labeling approach, which allows the use of a training set of categorized documents in one language to train a classifier for another language when an appropriate set of parallel documents is available.
- We evaluated the approach of combining categorization-based and corpus-based CLIR approaches.

We found that our categorization-based interlingual approach (using kNN) significantly outperforms the MT-based approach (using SYSTRAN), but it underperforms the corpus-based methods we examined. We obtained significant performance improvement by combining the categorization and corpus-based approaches versus using either alone.

In this paper, we focus on the medical domain. Extrapolating our work to diverse, multi-domain corpora and classification schemas is left to future work. Additionally, we would like to focus on effective use of domain-specific hierarchies of categories, by exploiting the hierarchical structure of the categories and automatically extract keywords per category for query expansion at different levels of granularity, and with user interaction. We also would like to improve classification performance by using Support Vector Machines or ridge regression, both of which have had better performance than kNN in recent evaluations [15][16].

Acknowledgements

We would like to thank Ralf Brown for collecting the SPRINGER data. This research is sponsored in part by the National Science Foundation (NSF) under grant IIS-9982226, and in part by the DOD under award 114008-N66001992891808. Any opinions and conclusions in this paper are the authors' and do not necessarily reflect those of the sponsors.

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] Jaime G. Carbonell, Yiming Yang, Robert E. Frederking, Ralf D. Brown, Yibing. Geng, D. Lee: Translingual Information Retrieval: A Comparative Evaluation. IJCAI 1997: 708-715

[3] David Eichmann, Miguel E. Ruiz and Padmini Srinivasan. Cross-language information retrieval with the UMLS Metathesaurus. In ACM SIGIR, pp. 72--80, 1998.

[4] Martin Franz, J. Scott McCarley, and Salim Roukos. Ad hoc and multilingual information retrieval at IBM. In The Seventh Text REtrieval Conference, pages 157--168, November 1998.

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

[6] Fraser, A., Xu, J., Weischedel, R. TREC 2002 Cross-lingual Retrieval at BBN. TREC 2002

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

[8] Gonzalo, J., F. Verdejo and I. Chugur (1999). Using EuroWordNet in a Concept-Based Approach to Cross-Language Text Retrieval. *Applied Artificial Intelligence 13(7)*,

[9] William R. Hersh, Chris Buckley, T. J. Leone, D. H. Hickam: OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research. SIGIR 1994: 192-201

[10] Hersh, W.R. and Donohoe, L. C. SAPHIRE International: a tool for cross-language information retrieval. Proceedingd/ AMIA Annual Symposium 1998: 673-7

[11] Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In Machine Learning: ECML-98.

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

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

[14] David D. Lewis, Robert E. Schapire, James P. Callan, Ron Papka: Training Algorithms for Linear Text Classifiers. SIGIR 1996: 298-306

[15] David D. Lewis, Yiming Yang, Tony Rose and Fan Li. RCV1: A new benchmark collection for text categorization research. Journal of Machine Learning Research (To appear) .

[16] Fan Li, Yiming Yang: A Loss Function Analysis for Classification Methods in Text Categorization. ICML 2003

[17] Oard, D. W. and F. Gey, "The TREC-2002 Arabic/English CLIR Track," TREC 2002

[18] Och, F. J. and Hermann N. Improved Statistical Alignment Models. In ACL 2000.

[19] Ogilvie, P and Callan, J. Experiments using the Lemur toolkit. In *TREC-10*. (2001)

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

[21] J. Rocchio. Relevance feedback information retrieval. 1971. In Gerard Salton, editor, The Smart retrieval system| experiments in automatic document processing, p 313-323.

[22] Monica Rogati and Yiming Yang. Multilingual Information Retrieval using Open, Transparent Resources in CLEF 2003 . In C. Peters(Ed.), Results of CLEF2003

[23] Monica Rogati and Yiming Yang. Resource Selection for Domain Specific CLIR . SIGIR 2004

[24] 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

[25] Volk, M., Ripplinger, B, Vintar, S, Buitelaar, P. Raileanu, D ; Sacaleanu, Bogdan: Semantic Annotation for Concept-Based Cross-Language Medical Information Retrieval. In: International Journal of Medical Informatics, Volume 67:1-3, December 2002.

[26] Yang, Y., Carbonell, J. G., Brown, R. and Frederking, R. E. Translingual Information Retrieval: Learning from Bilingual Corpora. Artificial Intelligence Journal: Best of IJCAI-97

[27] Yiming Yang, Xin Liu. A re-examination of text categorization methods. SIGIR 1999: 42-49