Carol Peters
Martin Braschler
Julio Gonzalo
Michael Kluck (Eds.)

LNCS 2785

Advances in Cross-Language Information Retrieval

**Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002
Rome, Italy, September 2002
Revised Papers**

CLEF

Springer

Lecture Notes in Computer Science

The LNCS series reports state-of-the-art results in computer science research, development, and education, at a high level and in both printed and electronic form. Enjoying tight cooperation with the R&D community, with numerous individuals, as well as with prestigious organizations and societies, LNCS has grown into the most comprehensive computer science research forum available.

The scope of LNCS, including its subseries LNAI, spans the whole range of computer science and information technology including interdisciplinary topics in a variety of application fields. The type of material published traditionally includes

- proceedings (published in time for the respective conference)
- post-proceedings (consisting of thoroughly revised final full papers)
- research monographs (which may be based on outstanding PhD work, research projects, technical reports, etc.)

More recently, several color-cover sublines have been added featuring, beyond a collection of papers, various added-value components; these sublines include

- tutorials (textbook-like monographs or collections of lectures given at advanced courses)
- state-of-the-art surveys (offering complete and mediated coverage of a topic)
- hot topics (introducing emergent topics to the broader community)

In parallel to the printed book, each new volume is published electronically in LNCS Online at *http://www.springerlink.com/series/lncs/*.

Detailed information on LNCS can be found at the series home page *http://www.springer.de/comp/lncs/*.

Proposals for publication should be sent to

LNCS Editorial, Tiergartenstr. 17, 69121 Heidelberg, Germany

E-mail: lncs@springer.de

ISSN 0302-9743

ISBN 3-540-40830-4

9 783540 408307

Lecture Notes in Computer Science

LNCS LNAI LNBI

springeronline.com

Table of Contents

Monolingual Experiments

Mainly Domain-Specific Information Retrieval

Interactive Track

Cross-Language Spoken Document Retrieval

II Cross-Language Systems Evaluation Initiatives, Issues and Results

III Appendix

luation Initiatives,

EC: A Status Report. In *Proceedings of th*
TREC-1), Gaithersburg, Maryland, Novem
52.
NTU at NTCIR3 MLIR Task. In *Workin*
o, October, 2002. National Institute of In
A Survey of Multilingual Text Retrieva
-19, University of Maryland, Institute fo
F-2001 Experiments: Effective Combine
Evaluation of Cross-Language Informatio
Computer Science, Vol. 2406, Darmstad
er, 27-43.
Johnson-Laird, B., 1995. The Collectio
of the Third Text REtrieval Conferen
November, 1994. NIST Publication, 9

SINAI at CLEF 2002: Experiments with Merging Strategies

Fernando Martínez, L. Alfonso Ureña, and M. Teresa Martín

Dpto. Computer Science
University of Jaén, Avda, Madrid 35, 23071 Jaén, Spain
{dofer,laurena,maite}@ujaen.es

Abstract. For our first participation in the CLEF multilingual task, we present a new approach to obtain a single list of relevant documents for CLIR systems based on query translation. This new approach, which we call two-step RSV, is based on the re-indexing of the retrieval documents according to the query vocabulary, and it performs noticeably better than traditional methods[1].

Introduction

usual approach in CLIR is to translate the query to each language present the corpus, and then run a monolingual query in each language. It is then necessary to obtain a single ranking of documents merging the individual lists from the separate retrieved documents. However, a problem is how to carry out such a merge? This is known as the merging strategies problem and is not trivial problem, since the weight assigned to each document (Retrieval Status Value - RSV) is calculated not only according to the relevance of the document and the IR model used, but also with respect to the rest of the monolingual corpus to which the document belongs [1].

There are various approaches to standardise the RSV, but in all cases a large decrease of precision is generated in the process (depending on the collection, between 20% and 40%) [2, 3]. Perhaps for this reason, CLIR systems based on document translation tend to obtain results which are noticeably better than those which only translate the query.

The rest of the paper is organized as follows. Firstly, we present a brief revision of the most extended methods for merging strategies. Sections 3 and 4 describe our proposed method. In Section 5, we detail the experiments carried out with the results obtained. Finally, we present our conclusions and future lines of work.

2 A Brief Review of the Merging Strategies

For each N language, we have N different lists of relevant documents, each obtained independently of the others. The problem is that it is necessary to obtain

1 This work has been supported by the Spanish Government (MCyT) with grant FIT-150500-2002-416.

C.A. Peters (Ed.): CLEF 2002, LNCS 2785, pp. 187–196, 2003.

a single list by merging all the relevant languages. If we suppose that each r trieved document of each list has the same probability to be relevant and th similarity values are therefore directly comparable, then an immediate approac would be simply to order the documents according to their RSV (this metho is known as raw scoring) [4, 5]. However, this method is not adequate, since th document scores computed for each language are not comparable. For exam ple, a document in Spanish that includes the term "información", can calculat a radically different RSV from another document in English with the same term "information". In general, this is due to the fact that the different indexing tech niques take into account not only the term frequency in the document (*tf*),but also consider how frequent such a term is in the rest of the documents, that is, the inverse document frequency (*idf*) [6]. Thus, the *idf* depends on each particular monolingual collection. A first attempt to make these values comparable is to standardise in some way the RSV of each document:

- By dividing each RSV by the maximum RSV obtained in each collection:

$$RSV_i' = \frac{RSV_i}{\max(RSV)}, 1 <= i <= N$$

- A variant of the previous method is to divide each RSV by the difference between the maximum and minimum document score values obtained in each collection [7]:

$$RSV_i' = \frac{RSV_i - \min(RSV)}{\max(RSV) - \min(RSV)}, 1 <= i <= N$$

in which RSV_i is the original retrieval status value, and $\max(RSV)$ and $\min(RSV)$ are the maximum and minimum document score values achieved by the first and last documents respectively. N is the number of documents in the collection.

However, the problem is only solved partially, since the normalization of the document score is accomplished independently of the other collections and, therefore, the differences in the RSV are still great.

Another approach is to apply a round-robin algorithm. In this case, the RSV obtained for each retrieved document is not taken into account, but rather the relative position reached by each document in their collection. A single list of documents is obtained and the document score m is in the position m in the list. Thus for example, if we have five languages and we retrieve five lists of documents, the first five documents of the single result list will coincide with the first document of each list; the next five, with the second document of each list; and so on. This approach is not completely satisfactory because the position reached by each document is calculated exclusively considering the documents of the monolingual collection to the one which belongs.

Finally, another approach, perhaps the most original, is to generate a single index with all the documents without taking into account the multilingual nature of the collection [8, 9, 10]. In this way, a single index is obtained in which the terms from each language are intermixed. In the same way as when all the documents in a single index are merged, we obtain a single query where the terms of several languages are also intermixed. That is, the query must be translated into each of the languages present in the multilingual collection. However, we do not generate a query for each translation, but merge all the translations forming a single query. This query will then be the one which we compare with the document collection. As with the approach based on document translation, in this approach the system will always return a single list of documents for each query. In spite of this, the problem is not eliminated: the ranking of each document is dependent on the language in which it is written. Although a single index is generated, the vocabulary of each language is practically exclusive. Two different languages rarely share terms. For this reason, the weight obtained by each term will refer to the language to which it belongs, and therefore, the similarity between documents will be correct with respect to the documents expressed in the same language. It should be mentioned that a notable exception is proper names, which are frequently invariable in different languages. In this case, this approach proves very effective.

A Useful Structure to Describe IR Models

In this section we present a notation that will be used to describe the proposed model. A large number of retrieval methods are based on this structure [11]:

$$< T, \Phi, D; ff, df >$$

where:

- D is the document collection to be indexed.
- Φ is the vocabulary used in the indices generated from D.
- T is the set of all tokens τ present in the collection D, commonly the words or terms. Thus, the function

$$\varphi : T \rightarrow \Phi, \tau \rightarrow \varphi(\tau)$$

  maps the set of all tokens, T,to the indexing vocabulary Φ. The function φ can be a simple process such as removing accents or another more complex process such as root extraction (stemming), lemmatization...
- ff is the feature frequency and denotes the number of occurrences of φ_i in a document d_j:

$$ff(\varphi_i, d_j) :=| \{\tau \in T \mid \varphi(\tau) = \varphi_i \wedge d(\tau) = d_j\} |$$

  where d is the function that makes each token τ correspond to its document:

$$d : T \rightarrow D, \tau \rightarrow d(\tau)$$

- df is the document frequency and denotes the number of documents containing the feature φ_i at least once:

$$df(\varphi_i) :=| \{d_j \in D \mid \exists \tau \in T : \varphi(\tau) = \varphi_i \wedge d(\tau) = d_j\} |$$

4 Two-Step Retrieval Status Value

The proposed method [12] is a system based on query translation and it calculates RSV in two phases, a pre-selection phase and a re-indexing phase. Although the method is independent of the translation technique, it is necessary to know how each term translates.

1. The document pre-selection phase consists of translating and running the query on each monolingual collection, D_i, as is usual in CLIR systems based on query translation. This phase produces two results:
   - we obtain a single multilingual collection of preselected documents (D' collection) as a result of joining all retrieved documents for each language.
   - we obtain the translation to the other languages for each term of the original query as a result of the translation process. That is, we obtain a T' vocabulary, where each element τ is called "concept" and consists of each term together with its corresponding translation. Thus, a concept is a set of terms expressed independently of the language.
2. The re-indexing phase consists of re-indexing the multilingual collection D', but considering solely the T' vocabulary. That is, only the concepts are re-indexed. Finally, a new query formed by the concepts in T' is generated and this query is executed against the new index. Thus, for example, if we have two languages, Spanish and English, and the term "casa" is in the original query and is translated by "house", both terms represent exactly the same concept. If "casa" occurs a total of 100 times in the Spanish collection, and "house" occurs a total of 150 times in the English collection, then the term frequency would be 250. From a practical point of view, in this second phase each occurrence of "casa" is treated exactly as each occurrence of "house".

Formally, the method can be described as follows:
For each monolingual collection we begin with the already-known structure:

$$< T_i, \Phi_i, D_i, ff, df >, 1 <= i <= N$$

Where N is the number of languages present in the multilingual collection to be indexed. Let $Q = \{Q_i, 1 <= i <= N\}$ be the set formed by the original query together with its translation into the other languages, in such a way that Q_i is the query expressed in the same language as the collection D_i. After each translation Q_i has been run against its corresponding structure $< T_i, \Phi_i, D_i, ff, df >$, it is possible to obtain a new and single structure:

$$< T', \Phi', D, D', ff', df' >$$

where:

- D is the complete multilingual document collection: $D = \{D_i, 1 <= i <= N\}$.

Value

ased on query translation and it calcu
hase and a re-indexing phase. Althoug
tion technique, it is necessary to kno

onsists of translating and running th
D_i, as is usual in CLIR systems base
duces two results:
ollection of preselected documents (D
all retrieved documents for each lan

other languages for each term of th
ranslation process. That is, we obtai
nt τ is called "concept" and consists o
sponding translation. Thus, a concep
ndently of the language.
ndexing the multilingual collection D
ary. That is, only the concepts are r
by the concepts in T' is generated an
v index. Thus, for example, if we hav
and the term "casa" is in the origin
oth terms represent exactly the sam
0 times in the Spanish collection, an
the English collection, then the term
ical point of view, in this second phas
xactly as each occurrence of "house"

d as follows:
gin with the already-known structure

$>, 1 <= i <= N$

resent in the multilingual collection t
e the set formed by the original quer
r languages, in such a way that Q_i is
as the collection D_i. After each trans
nding structure $< T_i, \Phi_i, D_i, ff, df >$
ructure:

$, ff', df' >$

ent collection: $D = \{D_i, 1 <= i <=$

- D' is the set of multilingual documents retrieved inas consequence of running the query Q.
- T' is the set of concepts τ_j, and denotes the vocabulary of the D' collection. Since each query Q_i is a translation of another, it is possible to align the queries at term level.

$$\tau_j := \{\tau_{ij} \in Q_i, 1 <= i <= N\}, 1 <= j = M, M = |Q|$$

where τ_{ij} represents all the translations of the term j of the query Q to the language i. Thus, τ_j denotes the concept j of the query Q independently of the language.
- Φ' is a new vocabulary to be indexed, such that each $\varphi_j \in \Phi'$ is generated as follows:

$$\varphi_j := \{\varphi(\tau_{ij}), 1 <= i <= N\}, 1 <= j <= M$$

The ff' function and df' function are interpreted as usual:

- ff' is the number of occurrences of the concept j in the document k. That is, the sum of the occurrences of the term j in the query, expressed in language i:

$$ff'(\varphi_j, d_k) := ff(\varphi_{ij}, d_k)$$

- df' is the number of documents with the concept j in the collection D. That is, the sum of the documents with the term j in the query, expressed in language i:

$$df'(\varphi_j) :=| \{d_k \in D_i \mid \exists \tau \in T : \varphi(\tau) = \varphi_j \wedge d(\tau) = d_k\} |$$
$$:= \Sigma df(\varphi_{ij}), \forall \varphi_{ij} \in \varphi_j, d_k \in D, 1 <= i <= N$$

where $df(\varphi_{ij})$ is all the documents that contain the concept j in the monolingual collection D_i.

Given this structure, a new index is generated in run time, but only taking into account the documents that are found in D'. The df function operates on the whole collection D, not only on the retrieved documents in the first phase, D'. This is because, in practice, we have found that the results obtained were slightly better when the whole collection was considered when calculating the idf factor. Once the indices have been generated in this way, the query Q formed by concepts, not by terms, is re-run on the D' collection.

In some ways, this method shares some ideas with CLIR systems based on corpus translation, but instead of translating the complete corpus, it only translates the words that appear in the query and the retrieved documents. These two simplifications allow the development of the system in run-query time since the necessary re-indexing process in the second phase is computationally possible due to the small size of the collection D' and to the scarce vocabulary T' (approximately, the query terms multiplied by the number of present languages in D').

Some relevant aspects of two-step RSV are:

- It is easily scalable to several languages.
- The system requires the term-level alignment of the original query and the translation of its terms. Depending on the approach followed for the translation, this process varies in complexity.
- A term together with its translation are treated in exactly the same way in the proposed model. This is not too realistic since it is not usual for the source term and its translations to be equally weighted. For example, it is possible that for a given language i, we maintain more than one translation for a given concept of the original query. Consequently, the concept frequency will be increased artificially in the documents expressed in the i language. In this case, if we know the translation probability of each term, we can weight each term according to its translation probability with respect to the source term. This can be modelled as follows:

$$ff'(\varphi_j, d_k) := \Sigma ff(\varphi_{ij}, d_k) * w(\tau_{ij}), \forall \varphi_{ij} \in \varphi_j, \varphi(\tau_{ij}) = \varphi_{ij}, 1 <= i <= N$$

where $w(\tau_{ij})$ represents the translation probability of each translation of term j in the query to language i, by default it will be 1.

5 Description of Experiments and Results

5.1 Multilingual Experiments

The experiment has been carried out for the five languages of the multilingual task. Each collection has been pre-processed as usual, using the stopword lists and stemming algorithms available for the participants, with the exception of Spanish, where we have used a stemming algorithm provided by the ZPrise system [2]. We have added terms such as "retrieval", "documents", "relevant"... to the stopword lists. Due to the morphological wealth of German, compound words have been reduced to simple words using the MORPHIX package [13]. Once the collections have been pre-processed, they are indexed with the Zprise IR system, using the OKAPI probabilistic model [14]. This OKAPI model has also been used for the on-line re-indexing process required by the calculation of two-step RSV.

For each query, we have used the Title and Description sections. The method for query translation is very simple: we used the Babylon[3] electronic dictionary to translate query terms [15]. For each term, we considered the first two translations given by Babylon. Words not found in the dictionary were been translated. This approach allows us to carry out query alignment at term level easily.

The results obtained show that the calculation of the two-step RSV improves more than seven points (36% more) the precision reached with respect to other approaches (Table 2). This improvement is approximately constant with short, medium and large queries (Table 3).

[2] ZPrise, developed by Darrin Dimmick (NIST). Available on demand at http://www.itl.nist.gov/iaui/894.02/works/papers/zp2/zp2.html

[3] Babylon is available at http://www.babylon.com

Table 1. Description of official experiments

Experiment	Task	Form	Query	Merging Strategy
UJAMLTDRR	Multilingual	automatic	Title+Description	Round-Robin
UJAMLTDNORM	Multilingual	automatic	Title+Description	Normalized score
UJAMLTDRSV2	Multilingual	automatic	Title+Description	2-Step RSV
UJAMLTDRSV2RR	Multilingual	automatic	Title+Description	2-Step RSV+ Round-Robin
UJABITD {SP,DE,FR,IT}	Bilingual	automatic	Title+Description	

Table 2. Performance using different merging strategies (official runs)

Experiment	Avg. prec.	R-Precision	Overall Recall
UJAMLTDRR	0.2038	0.2787	4246/8068
UJAMLTDNORM	0.2068	0.2647	4297/8068
UJAMLTDRSV2	0.2774	0.3280	4551/8068

Table 3. Average precision using different merging strategies and query lengths

Merging strategy	Tit.	Tit.+Desc.	Tit.+Desc.+Narr.
round-robin	0.1593	0.2038	0.2425
normalized score	0.1592	0.2068	0.2554
2-step RSV	0.2159	0.2774	0.3209

Table 4. Bilingual experiments (Title+Description)

Experiment	Language	Avg. prec.	R-Precision
UJABITDSP	english → spanish	0.2991	0.3141
UJABITDDE	english → german	0.2747	0.3077
UJABITDFR	english → french	0.3467	0.3365
UJABITDIT	english → italian	0.2438	0.2620

5.2 Bilingual Experiments

The differences in accuracy between the bilingual experiments may be due to the stemming algorithms used, the quality of which varies according to language. The simplest stemming algorithm is that used for Italian: it removes only inflectional suffixes such as singular and plural word forms or feminine and masculine forms, and it is in this language where the lowest level of accuracy is achieved.

Note that the multilingual document list has been calculated starting from the document lists obtained in the bilingual experiments. The accuracy obtained in the UJAMLTDRSV2 experiment is similar to that obtained in the bilingual

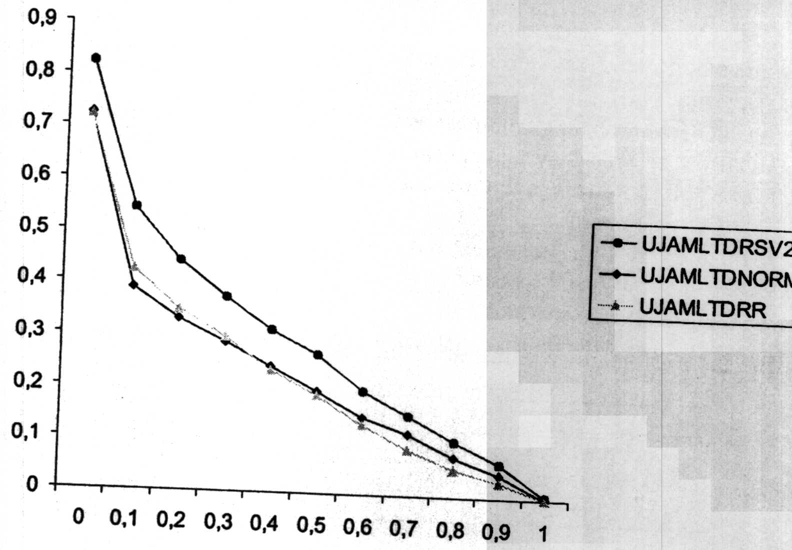


Fig. 1. 11 pt-precision

experiments (Table 4), surpassing even the accuracy for German and Ita and only two points short of that reached in Spanish.

5.3 Merging Several Approaches

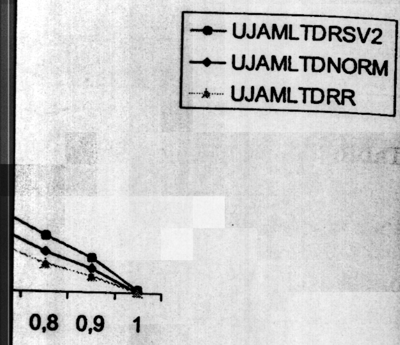
Finally, we carried out an experiment merging several approaches through a ple linear function. We calculated document relevance with the function:

$$Pos_i' = 0.6 * Pos_i^{rsv2} + 0.4 * Pos_i^{merge-approach}$$

Where Pos_i' is the new document position i. Pos_i^{rsv2} is the document posi reached using two-step RSV, and $Pos_i^{merge-approach}$ is the document posi using the Round-Robin or normalized score approach. As shown in Table 5, only is there no improvement, but the accuracy even decreases slightly.

Table 5. Merge of two-step RSV and round-robin/normalized score (tle+Description)

Experiment	Merging strategies	Avg. prec.	R-Precision
UJAMLTDRSV2	RSV2	0.2774	0.3280
UJAMLTDRSV2RR	RSV2 and round-robin	0.2758	0.3265
ujamltdrsv2norm	RSV2 and normalized score	0.2631	0.3162



-precision

he accuracy for German and Italia
in Spanish.

ing several approaches through a sim
nt relevance with the function:

$* Pos_i^{merge-approach}$

i. Pos_i^{rsv2} is the document positio
ge-approach is the document positio
e approach. As shown in Table 5, no
uracy even decreases slightly.

round-robin/normalized score (T

egies	Avg. prec.	R-Precision
	0.2774	0.3280
d-robin	0.2758	0.3265
ized score	0.2631	0.3162

6 Future Work

We have presented a new approach to solve the problem of merging relevant documents in CLIR systems. This approach has performed noticeably better than other traditional approaches. To achieve this performance, it is necessary to align the query with its respective translations at term level. Our next efforts are directed towards three aspects:

- We suspect that with the inclusion of more languages, the proposed method will perform better than other approaches. Our objective is therefore to confirm this suspicion.
- We intend to test the method with other translation strategies. We have a special interest in the Multilingual Similarity Thesaurus, since this provides a measure of the semantic proximity of two terms. This semantic proximity can be used by our method as the translation probability of a term.
- Finally, we could study the effect of pseudo-relevance feedback in the first and second phase of the method proposed.

References

[1] S. T. Dumais. Latent Semantic Indexing (LSI) and TREC-2. In *Proceedings of TREC-2*, pages 105–115, Gaithersburg, 1994.
[2] E. Voorhees. The collection fusion problem. In *Proceedings of TREC-3*, number 225, pages 95–104, Gaithersburg, 1995.
[3] J. Savoy. Report on CLEF-2001 Experiments. In Carol Peters, editor, *Proceedings of the CLEF 2001 Cross-Language Text Retrieval System Evaluation Campaign. Lecture Notes in Computer Science*, pages 10–19. Springer Verlag, 2001.
[4] K. L. Kwok, L. Grunfeld, and D. D. Lewis. TREC-3 ad-hoc, routing retrieval and thresholding experiments using PIRCS. In *Proceedings of TREC-3*, number 215, pages 247–255, Gaithersburg, 1995.
[5] A. Moffat and J. Zobel. Information retrieval systems for large document collections. In *Proceedings of TREC-3*, number 225, pages 85–93, Gaithersburg, 1995.
[6] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, London, U. K., 1983.
[7] A. L. Powell, J. C. French, J. Callan, M. Connell, and C. L. Viles. The impact of database selection on distributed searching. In The ACM Press., editor, *Proceedings of the 23rd International Conference of the ACM-SIGIR'2000*, pages 232–239, New York, 2000.
[8] A. Chen. Multilingual Information Retrieval Using English and Chinese Queries. In Carol Peters, editor, *Proceedings of the CLEF 2001 Cross-Language Text Retrieval System Evaluation Campaign. Lecture Notes in Computer Science*. Springer Verlag, 2002.
[9] F. Gey, H. Jiang, A. Chen, and R. Larson. Manual Queries and Machine Translation in Cross-language Retrieval and Interactive Retrieval with Cheshire II at TREC-7. In E. M. Voorhees and D. K. Harman, editors, *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pages 527–540, 2000.

[10] P. McNamee and J. Mayfield. JHU/APL Experiments at CLEF: Translation Resources and Score Normalization. In Carol Peters, editor, *Proceedings of the CLEF 2001 Cross-Language Text Retrieval System Evaluation Campaign. Lecture Notes in Computer Science.* Springer-Verlag, 2001.

[11] P. Sheridan, P. Braschler, and P. Schäuble. Cross-language information retrieval in a multilingual legal domain. In *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*, pages 253–268, 1997.

[12] F. Martínez-Santiago and L. A. Ureña. Proposal for an independient-language CLIR system. In *JOTRI'2002*, pages 141–148, 2002.

[13] G. Neumann. Morphix software package. http://www.dfki.de/~neumann/morphix/morphix.html.

[14] S. E Robertson, S. Walker., and M. Beaulieu. Experimentation as a way of life:okapi at trec. *Information Processing and Management*, 1(36):95–108, 2000.

[15] D. A. Hull and G. Grefenstette. Querying across languages. a dictionary-based approach to multilingual information retrieval. In *Procedings of 19th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–57, 1996.