Published in Information Retrieval 7, issues 1-2, 121-148, 2004
which should be used for any reference to this work

Combining Multiple Strategies for Effective Monolingual and Cross-Language Retrieval

JACQUES SAVOY

Institut interfacultaire d'informatique, Université de Neuchâtel, Switzerland
Jacques.Savoy@unine.ch

Abstract. This paper describes and evaluates different retrieval strategies that are useful for search operations on document collections written in various European languages, namely French, Italian, Spanish and German. We also suggest and evaluate different query translation schemes based on freely available translation resources. In order to cross language barriers, we propose a combined query translation approach that has resulted in interesting retrieval effectiveness. Finally, we suggest a collection merging strategy based on logistic regression that tends to perform better than other merging approaches.

Keywords: cross-language information retrieval, bilingual information retrieval, collection merging strategies, evaluation, learning curve

1. Introduction

The Cross-Language Evaluation Forum (CLEF) (Braschler and Peters, this volume) was founded to promote, study and evaluate information access technologies using various European languages. In this context, this paper presents the underlying problems encountered when implementing monolingual retrieval systems having to handle various non-English European languages. In fact, within the information retrieval (IR) domain, even though the language of Shakespeare has been studied for a relatively long period of time, there is currently a growing interest in other languages, including those with more complex morphologies than English.

In addition to the need to develop effective monolingual retrieval models, there is also an increasing need to promote bilingual retrieval systems able to accept queries expressed in one language in order to retrieve documents written in a different language. Finally, in multilingual countries such as Switzerland, or more generally in Europe, as well as in multinational companies or large international organizations, users would like to access multilingual information by submitting their requests to retrieval systems in their own language, even when searching for documents written in several other languages.

This paper will propose and evaluate various search strategies capable of working within monolingual, bilingual or multilingual contexts, based on the experience we gained during the exploratory Amaryllis cycle (Savoy 1999) and our participation in the Amaryllis evaluation campaign (Savoy 2002b). On the other hand, the indexing and search models suggested and described in this paper will be based on our participation in the CLEF 2001 (Savoy 2002a) and CLEF 2002 evaluation campaigns (Savoy 2003). In order to evaluate the

various search schemes presented in this article on a common basis, we will use the CLEF 2002 test collections.

This paper is organized as follows: Section 2 describes the progress made with monolingual IR systems when handling document collections written in French, Italian, German and Spanish. Section 3 evaluates several approaches used to resolve bilingual information retrieval problems, and finally Section 4 investigates and evaluates various merging strategies for multilingual systems in which corpora containing documents written in English, French, Italian, German and Spanish can be accessed through requests written in English.

2. Monolingual evaluation

Most European languages (including French, Italian, German, Spanish) share many characteristics belonging to the language of Shakespeare (e.g., word boundaries marked in a conventional manner, word variants generated by adding a suffix to the stem, etc.). Any adaptation of indexing or search strategies thus means the creation of general stopword lists and fast stemming procedures that can be used with other European languages. Stopword lists contain non-significant words that are removed from a document or a search request before the indexing process begins. Stemming procedures try to remove inflectional and derivational suffixes in order to conflate word variants into the same stem or root. In attempting to resolve these problems, it is important to remember that most European languages involve more complex morphologies than does the English language (Sproat 1992).

This section will deal with some of these issues, and is organized as follows: Section 2.1 contains an overview of the CLEF 2002 test collections and Section 2.2 describes our general approach to building stopword lists and stemmers for use with languages other than English. Section 2.3 describes our evaluation methodology. Section 2.4 depicts the various vector space term weighting schemes used in this paper together with the Okapi probabilistic model, and evaluates them using test collections and queries written in French, Italian, German and Spanish. Section 2.5 describes how we decompounded German words while Section 2.6 evaluates various combinations of document representations used to improve retrieval effectiveness when working with agglutinative languages such as German, Dutch or Finnish. Finally, Section 2.7 explains the learning curve resulting from our participation throughout the CLEF evaluation campaigns.

2.1. Overview of the test collections

The corpora used in this paper are those making up the CLEF 2002 test collections, extracted from newspapers such as the *Los Angeles Times* (1994, English) *Le Monde* (1994, French), *La Stampa* (1994, Italian), *Der Spiegel* (1994/95, German), and *Frankfurter Rundschau* (1994, German) together with various articles edited by news agencies such as *EFE* (1994, Spanish), and the Swiss news agency (1994, available in French, German and Italian but without parallel translation). For more information about CLEF see (Braschler and Peters, this volume). An examination of Table 1 reveals that the German and Spanish corpora included about twice as many articles as the collections for the other languages. Across

Table 1. Test collection statistics extracted from the CLEF 2002 test collection.

	English	French	Italian	German	Spanish
Size (in MB)	425	243	278	527	509
# of documents	113,005	87,191	108,578	225,371	215,738
# distinct terms	330,753	320,526	503,550	1,507,806	528,382
Number of distinct indexing terms/document					
Mean	167.33	130.213	129.908	119.072	111.803
Standard deviat.	126.315	109.151	97.602	109.727	55.397
Median	138	95	92	89	99
Maximum	1,812	1,622	1,394	2,420	642
Minimum	2	3	1	1	5
# of queries	42	50	49	50	50
# relevant items	821	1,383	1,072	1,938	2,854
Mean rel. items	19.548	27.66	21.878	38.76	57.08
Standard deviat.	20.832	34.293	19.897	31.744	67.066
Median	11.5	13.5	16	28	27
Maximum	96	177	86	119	321
Minimum	1	1	3	1	3

all corpora, the mean number of distinct indexing terms per document is relatively similar (around 120), although this number is a little bit higher for the English collection (167.33).

Within the CLEF 2002 test collection, there are 50 topics written in 12 different languages. Relevant documents can be found for these topics in most but not in all of the separate language collections. For the English and Italian corpora for example, relevant documents are found for 42 and 49 topics, respectively. Table 1 indicates the mean and median number of relevant documents per request found in this test collection. When computing the median for a distribution having an even number of observations, we return the mean of the middle two numbers (which for the English collection is a fractional number). The data in Table 1 reveals that the mean number is always greater than the median (e.g., for the French collection, there is an average of 27.66 relevant articles per query and the corresponding median is 13.5). The fact that the mean is greater than the median indicates that each collection contains numerous queries that retrieve a rather small number of relevant items or, in other words, the distribution of relevant items is positively skewed.

2.2. Stopword lists and stemming procedures

We defined a general stopword list containing many words determined to be of no use during retrieval, but found very frequently in document content. These stopword lists were developed for two main reasons: Firstly, we hoped that each match between a query and a document would be based only on pertinent indexing terms. Thus, retrieving a document

just because it contained words like “be”, “your” and “the” in the corresponding request does not constitute an intelligent search strategy. These non-significant words thus represent noise and actually damage retrieval performance, because they do not discriminate between relevant and irrelevant articles. Secondly, by using a stopword list, we can reduce the size of the inverted file, hopefully within the range of 30% to 50%. During our participation in the CLEF evaluation campaigns, we continually made efforts to enhance the stopword lists used for the various European languages (available at http://www.unine.ch/info/clef/).

Once the high frequency words have been removed, an indexing procedure uses a stemming algorithm in an attempt to conflate word variants into the same stem or root. In developing this procedure for the French, Italian, German and Spanish languages, our first attempt was to remove only inflectional suffixes such that singular and plural word forms or feminine and masculine forms would conflate to the same root. More sophisticated schemes have already been proposed for English, removing derivational suffixes (e.g., “-ably”, “-ship”, “-ize”). Examples are the stemmers developed by Lovins (1968) based on a list of over 260 suffixes, and Porter (1980), based on about 60 suffixes. In this vein, Figuerola et al. (2002) developed two different stemmers for the Spanish language, and the results show that removing only inflectional suffixes (88 different inflectional suffixes were defined) seemed to provide better retrieval levels than removing both inflectional and derivational suffixes (based on 230 suffixes).

Our various stemming procedures can be found at http://www.unine.ch/info/clef/. During the last CLEF evaluation campaign, we improved our stemming algorithm for the French language, removing some derivational suffixes. Finally, diacritic characters were replaced by their corresponding non-accentuated letter in the Italian, German and Spanish languages. Of course, other stemmers for various European languages have been suggested. Some examples include the Snowball string processing language at http://snowball.tartarus.org/ (MacFarlane 2003), the Xelda system at http://www.xrce.xerox.com/ats/xelda/, or statistical stemmers (Oard et al. 2001).

Given that French, Italian and Spanish morphologies are comparable to that of the English language, we decided to index French, Italian and Spanish documents based on word stems. For the German language and its more complex compounding morphology, we also decided to represent German articles and queries by using a 5-gram approach (McNamee and Mayfield 2002). However, contrary to McNamee and Mayfield (2002), our generation of 5-gram indexing terms does not span word boundaries. Using this indexing scheme, the compound ≪das Hausdach≫ (the roof of the house) will generate the following indexing terms: ≪das≫, ≪hausd≫, ≪ausda≫, ≪usdac≫ and ≪sdach≫. This value of 5 was chosen because it performed better with the CLEF 2001 corpora (Savoy 2002a).

2.3. Evaluation methodology

As a retrieval effectiveness indicator, we adopted the non-interpolated average precision (computed on the basis of 1,000 retrieved items per request by the TREC_EVAL program), thus allowing both precision and recall to be represented by a single number, as during the CLEF evaluation campaigns (Braschler and Peters 2002). To determine whether or not a given search strategy is better than another, a decision rule is required. To achieve this,

we could apply statistical inference methods such as Wilcoxon's signed rank test or the Sign test (Salton and McGill 1983, Section 5.2, Hull 1993). However, according to van Rijsbergen (1979), we know that the conditions required for the application of these tests are not really met in the information retrieval context.

> "The [Wilcoxon] test is done on the differences $D_i = Z_a(Q_i) - Z_b(Q_i)$, but it is assumed that D_i is continuous and that it is derived from a *symmetric* distribution, neither of which is normally met in IR data. It seems therefore that some of the more sophisticated statistical tests are inappropriate. ... It [sign test] makes *no* assumptions about the form of the underlying distribution. It does, however, assume that the data are derived from a *continuous* variable and that the $Z(Q_i)$ are *statistically independent*. These two conditions are unlikely to be met in a retrieval experiment. Nevertheless given that some of the conditions are not met it can be used *conservatively*" (van Rijsbergen 1979, pp. 178–179).

In order to overcome these difficulties, we based our statistical validation on the bootstrap methodology (Efron and Tibshirani 1993, Savoy 1997). This computer-based method can be used to assign an accuracy measure to virtually any statistical estimator. The basic idea of the bootstrap approach is simple and can be explained as follows. In order to measure retrieval effectiveness, we examine a sample of observations $X = \{x_1, x_2, \ldots, x_m\}$ of size m, drawn from a population having the probability distribution F. In our context, for query i, each x_i is the difference in average precision between situation a and situation b. If we know the real distribution F, we may compute the underlying parameter of interest, e.g., the mean, according to $\theta = t(F)$. Since the distribution F is unknown, we want to estimate the parameter θ using a point estimate $\hat{\theta} = t(\hat{F})$. This estimate will be computed according to the plug-in principle, whereby we use the same function, in our case $t()$, which should be applied if we know the real distribution F. In this computation, we substitute F by the empirical distribution $\hat{F}$.

The advantage of this bootstrap methodology is that the investigator does not have to make assumptions imposed by both parametric and non-parametric statistical models, or derive formulae that can be hard to come by. The bootstrap approach is however not an "assumption-free" method and requires that the observations are independent and identically distributed. In information retrieval, this means we must assume that the query samples associated with a given test collection are reasonable representatives of the request population.

In a statistical testing, the null hypothesis H_0 states that both retrieval schemes will result in similar average precision or $\text{mean}_a = \text{mean}_b$ (or $\text{mean}_a - \text{mean}_b = 0$ with a two-sided test, or $\text{mean}_a - \text{mean}_b \geq 0$ with a one-sided test). Such a null hypothesis plays the role of a devil's advocate, and this assumption will be accepted if the two retrieval schemes return statistically similar means, and will otherwise be rejected.

In the various tables found in this paper, we statistically analyzed the differences in average precision, based on a one-sided, non-parametric bootstrap test with a significance level fixed at 5%. This value of 5% means that when we decide to reject the null hypothesis, there is less than a 5% chance that H_0 is true, according to the observed values. Thus, if H_0 is rejected 100 times, there will be, in mean, 5 incorrect decisions (for 5 times we will reject H_0while $H_0$is true) due to random variability. On the basis of this observation, it is

important not to base a decision on only a single statistical test, and in this paper, our main conclusions will be based on a set of evidence. On the other hand, the decision to accept H_0 is not the equivalent of the null hypothesis H_0 being true, rather it represents the fact that "H_0 has not been shown to be false", resulting in insufficient evidence against H_0.

2.4. Indexing and searching strategies

In order to obtain a broader view of the relative merits of the various retrieval models used in the European languages, we first adopted a binary indexing scheme within which each document (or request) is represented by a set of keywords without any weight assigned. To measure the similarity between documents and requests, we counted the number of common terms, computed according to the inner product (retrieval model denoted "doc = bnn, query = bnn" or "bnn-bnn" using the terminology introduced by Salton and Buckley (1988)). For document and query indexing however, binary logical restrictions are often too limiting. In order to weight the presence of each indexing term in a document surrogate (or in a query), we could take into account term occurrence frequency (denoted tf) allowing for better term distinction and increasing indexing flexibility (retrieval model notation: "doc = nnn, query = nnn" or "nnn-nnn").

Terms in the collection that occur very frequently are not however considered very helpful in distinguishing between relevant and non-relevant items. We thus count their frequency in the collection (denoted df), or more precisely the inverse document frequency (denoted by $idf = \ln(n/df)$, with n indicating the number of documents in the collection), resulting in more weight for sparse words and less weight for more frequent ones. Moreover, a cosine normalization could prove beneficial and each indexing weight varies within the range of 0 to 1 (weighting scheme "doc = ntc, query = ntc"). Appendix 1 depicts the precise weighting schemes used in this paper.

Other variants might also be created, especially if we consider that the occurrence of a given term in a document is a rare event. Thus, a good practice is to give more importance to the first occurrence of this word, as opposed to any successive or repeating occurrences. The term frequency component may be therefore computed as $0.5 + 0.5 \cdot [tf/\max tf$ in a document] (term weighting scheme denoted "doc = atn"). Moreover, we should consider that a term's presence in a shorter document provides stronger evidence than it does in a longer document. To account for this, we integrated document length within the weighting formula, leading to more complex schemes; for example, the IR model denoted by "doc = Lnu" (Buckley et al. 1996), "doc = dtu" (Singhal et al. 1999).

In the vector space model, documents and queries are represented by vectors, while in the probabilistic model (Robertson and Sparck Jones 1976, van Rijsbergen 1979, Chapter 6), documents and requests representation, together with the decision to retrieve or not a given document, is based on probabilistic theory. Within this framework, various probabilistic models have been suggested, and in this paper, we will use the Okapi probabilistic model (Robertson et al. 2000). This retrieval model is based on the 2-Poisson model (Harter 1975), in which new variables such as term frequencies, document frequency and document length are incorporated in order to provide useful insights regarding the probability that a given document is relevant with respect to the request (Robertson and Walker 1994).

Table 2. Average precision of various IR models (based on CLEF 2002 test collection, monolingual).

	Average precision					
Query TD	French word 50 queries	Italian word 49 queries	Spanish word 50 queries	German word 50 queries	German decomp. 50 queries	German 5-gram 50 queries
doc = Okapi, que = npn	**48.41**	**41.05**	**51.71**	**37.39**	**37.75**	**39.83**
doc = Lnu, query = ltc	46.97	39.93	49.27	36.41	36.77	36.91
doc = dtu, query = dtc	45.38	39.53	47.29	35.55	35.08	36.03
doc = atn, query = ntc	42.42	39.08	46.01	34.48	33.46	37.90
doc = ltn, query = ntc	44.19	37.03	46.90	34.68	33.67	34.79
doc = ntc, query = ntc	31.41	29.32	33.05	29.57	31.16	32.52
doc = ltc, query = ltc	32.94	31.78	36.61	28.69	29.26	30.05
doc = lnc, query = ltc	33.49	32.79	38.78	29.33	29.14	29.95
doc = bnn, query = bnn	18.59	18.53	25.12	17.65	16.88	16.91
doc = nnn, query = nnn	14.97	15.63	22.22	14.87	12.52	8.94

Throughout this paper, in order to facilitate the reading of our evaluations, we have adopted the following typographical convention. The best performance for a given language or condition is always indicated in bold. Each statistically significant difference in average precision compared to a given baseline is underlined. If we need to compare a given approach with two different baselines, a statistically significant difference with these two baselines is denoted by double underlining.

The evaluation of various retrieval models based on queries using the Title and Description (denoted "TD") fields is reported in Table 2. Sometimes, we will also evaluate all topics fields, namely the Title, Description, and Narrative sections (denoted "TDN"). This data shows that the Okapi probabilistic model performs best with four different languages. Since this probabilistic approach consists of three parameters that must be fixed, the exact values attached to these parameters are depicted in Table 6. In the second position, we usually find the vector-space model "doc = Lnu, query = ltc" and in the third "doc = dtu, query = dtc". Finally, the traditional *tf-idf* weighting scheme ("doc = ntc, query = ntc") did not exhibit very satisfactory results, and the simple term-frequency weighting scheme ("doc = nnn, query = nnn") or the simple coordinate match ("doc = bnn, query = bnn") resulted in poor retrieval performance. However, Amati et al. (2003) indicated that the PROSIT probabilistic model performed better than the Okapi approach, at least for the Italian collection.

Based on the bootstrap hypothesis testing methodology, differences in average precision cannot always be viewed as significant (significance level of 5%) compared to the Okapi model. A closer look at Table 2 demonstrates that, for the French collection and when comparing the Okapi IR model with the "doc = dtu, query = dtc" vector-processing scheme, the mean difference was 6.3% (48.41 vs. 45.38) and in favor of the Okapi approach. The bootstrap test however did not detect any statistically significant difference, thus the performance value is not underlined. In our previous example, a query-by-query analysis revealed

that the Okapi probabilistic model improved retrieval effectiveness for 26 queries out of a total of 50. On the other hand, for 19 requests, the "doc = dtu, query = dtc" search scheme showed better retrieval performance, while for five requests, the average precision was the same. Thus, in order to find a statistically significant difference between the two retrieval schemes, the performance difference between individual requests should favor one given retrieval model for a large number of queries and also the difference must be significant (e.g., an improvement of 0.1% cannot be viewed as significant).

2.5. Decompounding German words

Many European languages manifest other morphological characteristics, where compound word constructions (e.g., newspaper, courtroom) are some of the most important ones to consider. Compound words are widely used in German and this causes more difficulties than it does in English. For example, a research project is "Forschungsprojekt" ("Forschung" + s + "Projekt" for research + project). The morphological marker ("s") is not always present, as for example in "Krankenhaus" (hospital) built as "Kranken" (sick person, patient) + "Haus" (house).

According to Monz and de Rijke (2002), including both compounds and their composite parts (only noun-noun decompositions in (Monz and de Rijke 2002)) in queries and documents can result in better search performance. However, according to Molina-Salgado et al. (2002), the decomposition of German words seems to reduce average precision. We also suggested an algorithm that splits compound German words into their components based on the application of linguistic rules used to build German compounds (Savoy 2003). As can be seen in Table 2, the retrieval performance of our decompounding approach listed under the label "decomp." is similar to that of a word-based indexing procedure.

As an alternative, we might also decompound German compounds using a list of German words in order to generate all possible ways of breaking down a compound and then selecting the decomposition having a minimal number of component words, as suggested by Chen (2002, 2003). Retaining the compounds and their component words in document representations but only the component words in the queries seems to be the most effective approach (Chen 2002). This matter has not however been satisfactorily resolved. For example, in his last paper, Chen (2003) suggested including only component words in both the document and request representations in order to obtain the best average precision. In our approach however, we suggest using a data fusion approach for the agglutinative languages, as will be described in the next section.

2.6. Data fusion

For the German language, our hypothesis involves the use of 5-gram indexing, decompound indexing and word based document representation methods as distinct and independent sources of evidence regarding the content of German language documents. We therefore decided to combine these three indexing schemes by applying various fusion operators, as suggested by Fox and Shaw (1994) and depicted in Table 3. For example, the combSUM operator indicates that the combined document score (or the final retrieval status value)

Table 3. Data fusion combination operators.

combMAX	MAX (RSV_i)
combMIN	MIN (RSV_i)
combSUM	SUM (RSV_i)
combANZ	SUM (RSV_i)/# of nonzero (RSV_i)
combNBZ	SUM (RSV_i) $*$ (# of nonzero (RSV_i))
combRSV%	SUM (RSV_i/MAX$_{RSV}$)
combRSVnorm	SUM [(RSV_i-MIN$_{RSV}$)/(MAX$_{RSV}$-MIN$_{RSV}$)]
CORI old	CORI in 1995 (Callan et al. 1995)
CORI new	CORI in 2000 (Callan 2000)

is simply the sum of the retrieval status value (RSV_i) as achieved by the three indexing schemes. CombNBZ specifies that we multiply the sum of the document scores by the number of retrieval schemes able to retrieve this given document. In this table, we can see that both the combRSV% and combRSVnorm apply a normalization procedure when combining document scores.

In addition to the data fusion operators suggested by Fox and Shaw (1994), we have also considered the round-robin approach whereby we take one document in turn from all individual lists and remove duplicates, keeping the most highly ranked instance. For the purpose of comparison, we also added two versions of the CORI models (Callan 2000) which are useful for combining the result lists supplied by different search systems (see Section 4.1).

Finally, we applied the logistic regression approach (Hosmer and Lemeshow 2000, Kleinbaum and Klein 2002) which predicts the probability of a binary outcome variable according to a set of explanatory variables. In our case, and based on previous work by Le Calvé and Savoy (2000), this fusion method can be used to predict the relevance probability of a given document, according to its retrieval status value and the natural logarithm of its rank. After estimating the relevance probability for each document, the corresponding probabilities were added if a given article was retrieved by more than one retrieval scheme. Instead of the original document score RSV_i, the resulting estimated probabilities (or the sum of them) was used when sorting the retrieved records, in order to obtain a single ranked list. However, to estimate the underlying parameters of the logistic regression, a training set is required. In our evaluation, this training set included all requests except the current query (the leaving-one-out evaluation strategy) which produced an unbiased estimator of the real performance of the evaluated data used in the fusion approach.

Table 4 displays an evaluation of these various data fusion operators compared to the single approaches using the Okapi model, for which the underlying parameters were fixed according to Table 6. As shown in Table 4, many fusion strategies improve the retrieval effectiveness. However, based on the bootstrap test (with a significance level of 5%), the improvement over the 5-gram indexing scheme is statistically significant only when applying the combSUM and the logistic regression approaches (values underlined in Table 4), where queries are built from the Title and Description sections ("TD") of the requests or from the Title, Description and Narrative logical sections ("TDN").

Table 4. Average precision of various data fusion strategies (based on CLEF 2002 test collection, German monolingual).

	Average precision (% change)	
	TD	TDN
Individual runs		
Okapi word	37.39	41.60
Okapi decompounding	37.75	41.67
Okapi 5-gram (baseline)	39.83	43.04
Combined runs		
Round-robin	40.18 (+0.9%)	44.02 (+2.3%)
combSUM	**42.31 (+6.2%)**	**46.70 (+8.5%)**
Logistic regression	41.97 (+5.4%)	45.88 (+6.6%)
combNBZ	41.49 (+4.2%)	45.92 (+6.7%)
CORI new	41.27 (+3.6%)	45.59 (+5.9%)
combRSVnorm	41.25 (+3.6%)	45.55 (+5.8%)
CORI old	40.63 (+2.0%)	45.16 (+4.9%)
combRSV%	40.59 (+1.9%)	45.15 (+4.9%)
combMAX	40.19 (+0.9%)	43.42 (+0.9%)
combANZ	28.82 (−27.6%)	29.41 (−31.7%)
combMIN	17.62 (−55.8%)	11.84 (−72.5%)

We also tried to apply various data fusion approaches when searching collections written in French, Italian and Spanish. For these languages, combining the word-based and 5-gram indexing schemes does appear to improve average precision, when compared to the single word-based indexing approach. For the Dutch and Finnish languages, we used the combRSVnorm operator to combine word-based and 5-gram document representations (Savoy 2003). For the Dutch language however, the combined model usually enhanced the retrieval performance while for the Finnish language it did not.

2.7. Monolingual IR learning curve

Sections 2.4 and 2.6 show the best indexing and searching approaches for various European languages. During our participation in the CLEF 2001 workshop, we were not able to achieve adequate performance levels for different reasons. Firstly, when faced with a new collection and a fortiori with a new language, we did not know which underlying Okapi model parameters would be best. Thus we applied a default parameter settings (avdl = 900, $b = 0.75$, and $k_1 = 1.2$) based on our pre-CLEF experience. Secondly, the stopword lists and stemming procedures we used in CLEF 2001 were relatively simple. For the CLEF 2002 campaign, we improved our French suffix-stripping algorithm so that it takes account of some derivational suffixes and for the French and German languages we enhanced the

Table 5. Comparison of performances based on last two CLEF monolingual experiments (Okapi model).

	Average precision (% change)				
Query TD	English 42 queries	French 50 queries	Italian 49 queries	Spanish 50 queries	German 5-gram 50 queries
CLEF01 (default)	48.63	43.51	40.50	50.22	39.47
CLEF02 (stemming)	48.63 (+0%)	47.12 (+8.3%)	40.50 (0%)	50.27 (+0.1%)	39.52 (+0.1%)
CLEF02 (optimum)	**50.08** (+3.0%)	**48.41** (+11.3%)	**41.05** (+1.4%)	**51.71** (+3.0%)	39.83 (+0.9%)
CLEF02 (combSUM)					**42.31** (+7.2%)

stopword lists. Finally, for the German corpus, we suggested using a data fusion approach based on the combSUM operator.

To analyze the relative merit of each of these modifications, in the first line of Table 5 we reported the average precision achieved by the search models presented at CLEF 2001 under the label "CLEF01 (default)" using the CLEF 2002 test collections. These performances were the result of applying the Okapi probabilistic model with default parameter setting and the first version of our stopword lists and stemmers. The second line (labeled "CLEF02 (stemming)") displays the performances achieved for the CLEF 2002 test collections after modifying our French stemmer and stopword lists (for the Spanish language, a few words were also added to the stopword list). The line starting with the label "CLEF02 (optimum)" indicates the average performance achieved when using the "optimum" parameters setting for the Okapi model, as depicted in Table 6. Thus in our case, varying the underlying Okapi model parameters does not really improve retrieval effectiveness. Finally for the German language and as depicted in the last line of Table 5, we adopted a data fusion approach that significantly improved retrieval effectiveness.

From the data shown in Table 5, we can see that adapting the underlying Okapi model parameters enhances retrieval effectiveness for all languages, but improvements made to the "CLEF01 (default)" do not produce statistically significant results for the Italian and German corpora. For the French language a more aggressive stemmer significantly enhanced average precision and for the German language, our data fusion approach seemed to be an appropriate

Table 6. Optimum parameters setting for the Okapi model.

Language	b	k_1	avdl
English	0.8	2	900
French	0.7	2	750
Italian	0.6	1.5	800
Spanish	0.5	1.2	300
German (word)	0.55	1.5	600
German (decomp.)	0.55	1.5	600
German (5-gram)	0.55	1.5	600

choice for handling an agglutinative language. Overall, improvements represented by our monolingual IR systems when comparing our participation in two CLEF experiments are clearly variable across these languages. For both the French and the German languages we clearly improved our IR models (+11.3% and +7.2% respectively) due to a new stemming algorithm as well as a combined indexing and search strategy. The problem yet to be solved here is whether or not an enhanced stemming approach will improve retrieval performance for the Italian and Spanish languages. For the English, Italian and Spanish languages, however, the simple adaptation of the underlying Okapi model parameters only marginally enhanced retrieval effectiveness (from +1.4% to +3.0%). By modifying the values of these parameters, Brand and Brünner (2003) were able to make more detailed evaluations of their effects on retrieval effectiveness, showing that an appropriate value for the parameter k_1 is around 1.6 and for b, the best value seems to be around 0.5. These authors also demonstrate that varying the value of b will have more impact on retrieval effectiveness.

3. Bilingual information retrieval

In the previous section, we obtained a better view of the progress made during the last CLEF evaluation campaigns concerning language-dependent retrieval approaches, showing that search models that perform well for English may also do so for other languages. For these languages that have compound constructions and for the Germanic family in particular, a data fusion approach combining two or more document representations may be able to produce better retrieval effectiveness.

In this section we will describe the underlying problems involved in effective bilingual IR, search systems that based on a query written in a given source language (English in our case), can retrieve relevant documents from a collection written in another target language. During the last two CLEF campaigns, in an attempt to cross these language barriers, we based our approach on freely and readily available translation resources. More precisely, we studied machine translation MT system that will automatically provide a complete translation of a given request into the desired target language, and also bilingual dictionary tools able to provide one or more translation alternatives for each search keyword. Choosing the English language for the request is not arbitrary, given that for this language there are a larger number of freely available translation resources. Moreover, for some specific language pairs, the single translation device available is usually a bilingual dictionary.

In Section 3.1, we describe some of the most effective bilingual systems suggested during the last CLEF evaluation campaigns. Section 3.2 presents our combined strategy and compares the retrieval effectiveness of our approach to other proposed solutions. In this section, we will also evaluate the progress made in this context.

3.1. Related work

During the first CLEF campaign, most participants chose to cross the language barrier by translating the queries into the target language. To achieve this, a large majority of the suggested approaches were based only on one translation resource, a bilingual dictionary in most cases. As an alternative, some participants proposed using either a MT system,

usually the SYSTRAN or the L&H PowerTranslator system, or an aligned parallel corpora (McNamee et al. 2001, Chen 2002, 2003). When such corpora were not available, some authors suggested building them using Web pages available in various languages (Nie et al. 1999, Hiemstra et al. 2001).

An analysis of MT translation system retrieval performance usually revealed more effective retrieval than did the aligned corpus approach (McNamee et al. 2001, Hiemstra et al. 2001). Moreover, the performance of parallel corpora usually did not prove to be very interesting in terms of overall retrieval effectiveness (Nie et al. 2001). As an explanation of this poor performance, various authors mention that the quality of sources (e.g., Web sites) and the size of available corpora are of prime importance (Nie et al. 2001, Braschler and Schäuble 2001, Braschler et al. 2002). Cultural, thematic and time differences may also play a role in the effectiveness of these approaches (Kwok et al. 2001). When using the appropriate aligned corpus however, it is possible to achieve good average precision levels, at least with German queries translated into English (McNamee and Mayfield 2002) or with English requests translated into French (Chen 2003).

In order to cross the language barriers, other approaches have been suggested, including that of Braschler and Schäuble (2001). They proposed building a similarity thesaurus taken from available and comparable corpora. These collections would then provide "pseudo-translation", meaning not a direct translation of the search keywords into the target language, but this approach provides a set of related terms in the target language, those most similar to the query viewed as a whole. Such an approach may work satisfactorily if the available corpora are of good quality and of a reasonable size. However, upon evaluating various translation strategies, these authors found that MT translation systems seemed to provide better retrieval performance levels than did similarity thesauri.

As a second general approach to promoting bilingual IR systems, computers might generate a unified collection by translating all documents into a common language (Braschler and Schäuble 2001). Although requiring extensive computation, this strategy may work better on a static collection of documents and it does not require a merging procedure (see Section 4).

As a third approach, several attempts have been made to combine various translation resources, for example two MT systems (Gey et al. 2001, Chen 2002, 2003) or both query translations with document translation strategies (Braschler and Schäuble 2001, Braschler et al. 2002). In this case, combining different translation resources usually produces a better performance than does a single translation approach.

In order to limit translation ambiguity, McNamee et al. (2001) or Ballesteros and Croft (1998) suggested adding terms to the submitted query before translating it into the target language. In this case, the query is used to search within a comparable collection of documents written in the request language and based on a pseudo-relevance feedback scheme, new and related terms being added to the query before translation.

3.2. Bilingual experiments

In our bilingual experiments, we were faced with the following situation. We used the English set of queries provided in the CLEF 2002 test collection but we did not have any

parallel or aligned corpora from which we could derive statistically or semantically related words in the target language (Nie and Simard 2002). In order to develop a fully automated approach, our first bilingual IR model translated the requests using the SYSTRAN™ (http://babel.altavista.com/translate.dyn) system (Gachot et al. 1988) or we translated search terms word-by-word using the BABYLON™ (http://www.babylon.com) bilingual dictionary. However, a bilingual dictionary might suggest not only one, but several candidates for each word, thus revealing the underlying ambiguity of a given term. In order to distinguish between different variants when looking at this bilingual dictionary, we took account only of the first translation alternative under "BABYLON 1", the first two translations under "BABYLON 2" and the first three translated terms under the label " BABYLON 3".

With the experience gained from our participation in CLEF 2001, we also examined other machine translations tools (namely, GOOGLE (http://www.google.com/language_tools), FREETRANSLATION (http://www.freetranslation.com), INTERTRAN (http://www.tranexp.com:2000/InterTran) and REVERSO (http://translation2.paralink.com)) during the CLEF 2002 evaluation campaign. Listed in Table 7 are the various retrieval performances obtained using different machine translation systems and the performance achieved by using the BABYLON bilingual dictionary. The performance achieved by the Okapi probabilistic model for human-based translation queries will constitute the baseline (row labeled "Human translation"). From our examination of four languages, and for the three different document representations for the German collection, we were able to observe that all translation approaches were characterized by an average precision statistically lower than the manually translated queries (bootstrap statistical testing method, significance level of 5%).

Finally, in the last line of Table 7 ("Best translation"), we report the average precision resulting from the best available translation on a per query basis. This must be viewed as a theoretical upper bound based on an oracle that always selects, without any error, the best translation for each request. Actually however, we do not know how to select this best

Table 7. Performance of various machine-based translation resources (based on CLEF 2002 test collections).

	Average precision					
Query TD Translation resource	French	Italian	Spanish	German word	German decomp.	German 5-gram
Human translation	48.41	41.05	51.71	37.39	37.75	39.83
SYSTRAN	42.70	32.30	38.49	28.75	28.66	27.74
GOOGLE	42.70	32.30	38.35	28.07	26.05	27.19
FREETRANSLATION	40.58	**32.71**	40.55	28.85	**31.42**	27.47
INTERTRAN	33.89	30.28	37.36	21.32	21.61	19.21
REVERSO	39.02	N/A	**43.28**	**30.71**	30.33	**28.71**
BABYLON 1	**43.24**	27.65	39.62	26.17	27.66	28.10
BABYLON 2	37.58	23.92	34.82	26.78	27.74	25.41
BABYLON 3	35.69	21.65	32.89	25.34	26.03	23.66
Best translation	51.29	40.47	51.11	39.58	41.13	39.33

translation, based either on statistical properties or on a linguistic-based model. Comparing these best translation performances with those achieved by the manually translated requests ("Human translation"), the difference in average precision is obviously rather small (statistically not significant), indicating that machine based query translation can be a valid approach.

Moreover, the best single translation system varies across languages. For example, while the REVERSO machine translation system appears to be the best approach for both the Spanish and two German representations, FREETRANSLATION is the best for Italian and the bilingual dictionary BABYLON for the French language. If we compare each translation tool with the best translation system for a given language, we usually discover that only a few resources perform at lower levels that can be viewed as statistically significant (double-underlined values in Table 7). For the Italian language for example, the best translation system was FREETRANSLATION and only the three bilingual dictionary approaches (namely "BABYLON 1", " BABYLON 2" and "BABYLON 3") produced performance levels statistically lower than this translation resource. For the moment, we cannot explain why a given translation resource might work well for a given language and poorly for another. Except for the French language however, the solution given by the REVERSO system usually produces a good translation.

We also know that each overall statistic, such as average precision, may hide performance irregularities among requests when comparing two retrieval schemes. In Table 8 we quantify this phenomenon reporting the best translation tool for each language and the number of requests for which this best translation system performs best (line labeled "Best result for # query"). From this data, we can see that the best approach provides more precise results in 11–14 queries out of 50. In this same table, we included a couple of statistics from each translation alternative. The first indicates the number of queries for which this alternative translation resource results in a 10% better average precision compared to the best translation device and in a second position the number of requests whose retrieval

Table 8. Query-by query-analysis of performance variation, compared to the best translation system.

	Number of queries					
Query TD Translation tools	French	Italian	Spanish	German word	German decomp.	German 5-gram
Best translation	BABYLON 1	FREE	REVERSO	REVERSO	FREE REVERSO	
Best result for # query	13	12	11	13	13	14
SYSTRAN	14/23	16/26	11/18	16/20	18/21	11/18
GOOGLE	14/23	16/26	10/18	13/17	14/17	9/17
FREETRANSLATION	17/23		12/20	13/18		12/16
INTERTRAN	10/13	14/20	16/22	10/14	12/14	9/11
REVERSO	11/22	N/A			23/29	
BABYLON 1		12/17	10/20	15/20	15/23	12/19
BABYLON 2	13/16	11/15	6/13	13/17	22/26	8/12
BABYLON 3	8/15	11/15	9/12	14/19	16/22	8/11

effectiveness is greater than the best approach. For the French language for example, the best translation tool was BABYLON 1, producing the best average precision for 13 requests. The REVERSO system however produced 22 requests performing better than BABYLON 1, in which 11 had an average precision greater than 10% over the corresponding BABYLON 1 translated request.

Based on this larger sample, this experiment confirmed previous studies demonstrating that, for a particular request, the best translation tool does not always produce the best translation (Braschler and Schäuble 2001, Braschler et al. 2002, Hiemstra et al. 2001, McNamee et al. 2001). Moreover, Table 8 shows that this best translation approach only provides the best translation for a rather small number of requests (11 to 14 over a total of 50).

This fact also confirms that when a translation resource misses a few important search keywords, the resulting performance is seriously affected. Therefore, a combination of translation resources will help remedy failures caused by individual translation systems.

In order to evaluate the progress made during the last three years, we had to chose an automatic translation strategy that represented the state of the art in CLEF 2000 because we did not participate in this evaluation campaign. In Table 9, the BABYLON bilingual dictionary represents this query translation strategy because it was the most popular approach used in the CLEF 2000 campaign. Thus, under the label "CLEF00: BABYLON 1" each search term is automatically translated by taking only the first translation alternative provided by this bilingual dictionary.

In our CLEF 2001 participation when translating the English requests, we suggested combining two translation resources. In this case, our automatically translated queries were composed of all the words translated by the bilingual dictionary and the translated sentence furnished by the SYSTRAN system (Gey et al. 2001, 2002, Chen 2002, 2003). As shown in Table 9 under the label "SYSTRAN+BAB1", this combined query translation approach improves average precision over any single translation scheme by about 7.8% for the French language to 18.2% for the Italian collection. Based on the bootstrap test, the difference between CLEF 2000 and CLEF 2001 query translation strategies was always significant and in favor of the CLEF 2001 approach (significance level of 5%). For the German language,

Table 9. Comparison of performance based on last three CLEF automatic query translation strategies.

	Average precision (% change)			
Query TD	French 50 queries	Italian 49 queries	Spanish 50 queries	German 5-gram 50 queries
CLEF00: BABYLON 1	43.24	27.65	39.62	28.10
CLEF01: SYSTRAN+BAB1	46.63 (+7.8%)	32.68 (+18.2%)	43.97 (+11.0%)	31.87 (+13.4%)
CLEF02: combRSVnorm	48.56 (+12.3%)	35.82 (+29.6%)	45.63 (+15.2%)	33.34 (+18.7%)
				38.71 (+37.8%)
Monolingual	48.41	41.05	51.71	42.31
CLEF02	48.56 (+0.3%)	35.82 (−12.7%)	45.63 (−11.8%)	38.71 (−8.5%)

we only considered the 5-gram approach because the other two approaches (word-based or our decompounding scheme) showed a similar pattern.

During the CLEF 2002 campaign, we enlarged our query translation approach. On the one hand, we knew that the best translation tools are language-dependent, as depicted in Table 7 and are characterized by a large variability for a given language (see Table 8). On the other hand, we considered combining more than two translation resources. In this vein, we combined the REVERSO, FREETRANSLATION, GOOGLE and BABYLON 1 translations when searching the French corpus, FREETRANSLATION and GOOGLE for Italian, REVERSO, SYSTRAN and BABYLON 1 for Spanish, and REVERSO, GOOGLE and BABYLON 1 for German. Other combinations of translation resources having a lower retrieval performance can however be found in Savoy (2003).

As shown in Table 9, this extended and language-dependant combined query translation approach improves the performance over both the CLEF 2000 and CLEF 2001 translation strategies. When using our bootstrap inference approach, we can conclude that the CLEF 2002 query translation approach is always significantly better than the CLEF 2000 translation scheme (values underlined in Table 9). However, this statistical test does not always reveal a significant difference when comparing CLEF 2001 and CLEF 2002 automatic translation strategies (values double underlined in Table 9). Finally, in examining our data fusion approach for the German corpus (performance depicted in the line labeled "combRSVnorm"), the retrieval effectiveness achieved is statistically better compared to both the CLEF 2000 and CLEF 2001 translation strategies (performance double underlined).

When examining the average precision resulting from the CLEF 2002 query translation strategies and the monolingual runs used as baselines (second part of Table 9), we can see that for the French language the performance difference is rather small and not significant. For the Italian and Spanish collections, the bilingual IR system produces an average precision of around 12% less than the monolingual model and these differences are statistically significant. For the German language, the bilingual IR system based on the data-fusion model results in lower performance levels but the difference cannot be viewed as significant.

4. Merging strategies for multilingual systems

The previous section showed the retrieval effectiveness of the bilingual retrieval systems proposed during the last three CLEF evaluation campaigns. In this section, we will describe the learning effects obtained as a result of proposing better merging strategies over the same time period. During these evaluation campaigns, the majority of effective multilingual information retrieval (MLIR) systems (Savoy 2002a, 2003, Kraaij 2002, Chen 2002, 2003) divided up the set of all available documents in accordance to document language. After automatically translating the request into the corresponding target language (see Section 3) and obtaining a result list for each language, an MLIR system needs to effectively merge the results and then to present a single list of retrieved articles to the users.

In this section, we will evaluate different merging strategies based on the following situation. The multilingual retrieval system received a request in English in order to retrieve relevant documents in English, French, German, Italian and Spanish. As described in the previous section, in order to effectively confront this multi-language barrier, the various

collections are indexed separately using a language specific procedure. Moreover, we used the set of requests provided in the CLEF 2002 test collection in order to evaluate which of the merging strategies suggested during the last three years are most effective across the various collections. Finally, in this section, we only considered the Okapi probabilistic scheme because this search model shows the best retrieval performance and also because our aim is to evaluate various merging strategies based on a good retrieval model.

Section 4.1 describes the strategies that were proposed or most used for merging during the last CLEF campaigns. Section 4.2 analyzes and evaluates the various merging approaches based on manually or automatically translated requests. Moreover, this section also provides a quantitative view of the progress made in this matter during the previous CLEF evaluation campaigns.

4.1. Related work

As a first approach towards merging various result lists provided by each collection or language, we might assume that each collection contains approximately the same number of pertinent items and that the distribution of relevant documents is similar across the result lists. Using the rank as the sole criteria, we can interleave the retrieved records in a round-robin fashion, a strategy used by various multilingual information retrieval systems in the first CLEF campaign (Braschler and Schäuble 2001). This merging strategy will be used as the baseline for comparisons in our evaluations (see Table 10(a)–(d)).

When using this merging strategy on documents written in the same language (e.g., English), previous studies (Voorhees et al. 1995, Callan et al. 1995) demonstrated that the retrieval effectiveness is below (−40%) that achieved from a single retrieval scheme, working with a single huge collection that represents the entire set of documents. However,

Table 10(a). Evaluation of various merging strategies (based on manually translated queries).

	Average precision (% change)				
Query TD Model	English word 42 queries	French word 50 queries	Italian word 49 queries	Spanish word 50 queries	German combSUM 50 queries
Okapi-npn	50.08	48.41	41.05	51.71	42.31
Merging strategy (50 queries)					Average precision (% change)
CLEF00: Round-robin (baseline)					33.85
CLEF00: Raw-score					12.53 (−63.0%)
CLEF01: combRSV%					35.09 (+3.7%)
CLEF01: combRSVnorm					36.45 (+7.7%)
CLEF02: Biased Round-robin					35.85 (+5.9%)
CLEF02: Logistic regression					**38.82 (+14.7%)**
CORI old					26.87 (−20.6%)
CORI new					36.41 (+7.6%)

Table 10(b). Evaluation of various merging strategies (based on manually translated queries and with query expansion).

	Average precision (% change)				
Query TD Model	English word 42 queries	French word 50 queries	Italian word 49 queries	Spanish word 50 queries	German combSUM 50 queries
Okapi-npn	50.08	53.18	46.35	56.95	46.73
# documents	0	5	5	10	5
# added terms	0	15	20	30	40
Merging strategy (50 queries)					Average precision (% change)
CLEF00: Round-robin (baseline)					37.28
CLEF00: Raw-score					14.33 (−61.6%)
CLEF01: combRSV%					36.71 (−1.5%)
CLEF01: combRSVnorm					38.40 (+3.0%)
CLEF02: Biased Round-robin					39.60 (+6.2%)
CLEF02: Logistic regression					**43.79 (+17.5%)**
CORI old					29.58 (−20.7%)
CORI new					38.14 (+2.3%)

Table 10(c). Evaluation of various merging strategies (based on machine based translated queries).

	Average precision (% change)				
Query TD Model	English word 42 queries	French word 50 queries	Italian word 49 queries	Spanish word 50 queries	German combRSVnor 50 queries
Okapi-npn	50.08	48.56	35.82	45.63	38.71
Merging strategy (50 queries)					Average precision (% change)
CLEF00: Round-robin (baseline)					31.16
CLEF00: Raw-score					16.07 (−48.4%)
CLEF01: combRSV%					31.81 (+2.1%)
CLEF01: combRSVnorm					34.04 (+9.2%)
CLEF02: Biased Round-robin					32.57 (+4.5%)
CLEF02: Logistic regression					**34.86 (+11.9%)**
CORI old					25.70 (−17.5%)
CORI new					34.03 (+9.2%)

Table 10(d). Evaluation of various merging strategies (based on machine based translated queries and with query expansion).

	Average precision (% change)				
Query TD Model	English word 42 queries	French word 50 queries	Italian word 49 queries	Spanish word 50 queries	German combRSVnor 50 queries
Okapi-npn	50.08	52.04	40.11	51.22	43.47
# documents	0	10	5	10	5
# added terms	0	15	30	100	125
Merging strategy (50 queries)					Average precision (% change)
CLEF00: Round-robin (baseline)					34.60
CLEF00: Raw-score					5.75 (−83.4%)
CLEF01: combRSV%					33.83 (−2.2%)
CLEF01: combRSVnorm					37.02 (+7.0%)
CLEF02: Biased Round-robin					34.80 (+0.6%)
CLEF02: Logistic regression					**39.78 (+15.0%)**
CORI old					27.65 (−20.1%)
CORI new					36.79 (+6.3%)

this difference in performance has been shown to diminish (around −20%) when considering another collection (Savoy and Rasolofo 2001).

In order to account for the document score computed for each retrieved item (or the similarity value between the retrieved record and the request), we might formulate the hypothesis that each collection is searched by the same or a very similar search engine and that the similarity values are therefore directly comparable (Kwok et al. 1995). Such a strategy, called raw-score merging, produces a final list sorted by the document score computed by each collection and it will be evaluated in the next section.

In the CLEF evaluation campaigns, various participants followed this assumption by performing searches in different languages using the same retrieval scheme and usually the same indexing procedure (Hiemstra et al. 2001). For example, Gey et al. (2001), Chen (2002, 2003) proposed merging result lists provided by collections searched using the same retrieval model. In order to obtain adequate retrieval performance levels, these authors had to correct the translated query term weights (since the translated requests included more than one translation source) and to increase the document score of the top-ranked 10 documents (or top 50 in Chen (2002)) for each collection to ensure that these top-ranked articles are included in the final result list (Chen 2003).

However, as demonstrated by Dumais (1994), collection-dependent statistics in document or query weights may vary widely among collections, and therefore this phenomenon may invalidate the raw-score merging hypothesis. But different evaluations carried out using

English only documents have demonstrated that the raw-score merging strategy sometimes leads to satisfactory performance (Savoy and Rasolofo, 2001).

As a third merging strategy, we normalized document scores within each collection through dividing them by the maximum score (i.e. the document score of the retrieved record in the first position (Fox and Shaw 1994), a strategy denoted "combRSV%") in order to obtain more comparable document scores across collections. As a variant of this normalized score merging scheme, Powell et al. (2000) suggested normalizing the document scores by taking the maximum and minimum document score into account, as shown in Table 3 by the formula "combRSVnorm". Other variants have been suggested as a means of obtaining satisfactory merging performances (Hiemstra et al. 2001). For example, McNamee and Mayfield (2002) suggested normalizing the document score of each individual article using the sum of the scores assigned to the top 1,000 retrieved items.

As a fourth merging strategy, Callan et al. (1995) suggested a merging strategy named CORI, based on the score achieved by both collection and document. The collection scores are computed according to the probability that the corresponding collection respond appropriately to the current request. The corresponding collection score will be used to modify the similarity value attached to each document. Instead of using this document score directly (as in the raw-score merging strategy), the final document score is equal to the collection weight multiplied by its original document score. However this first CORI merging strategy (denoted "CORI old") may not perform very well because document scores produced by each collection are based on different corpus statistics and possibly different representations, and therefore the resulting scores are not directly comparable. In order to eliminate the requirement for specific cooperation between servers or collections, Callan (2000) suggested adding an heuristic to the CORI model (denoted "CORI new"). This issue is important in our CLEF test collection because evaluations available have demonstrated that the *idf* score, and hence document scores, are highly skewed. For example, documents from a given collection having a search keyword in common (and thus a good collection for this search term) tend to have low scores due to low idf values while documents extracted from another collection where the same search term is rare tend to have high scores (due to high *idf* values).

As a fifth merging strategy, Savoy (2003) suggested a merging approach based on the logistic regression approach (see Section 2.6) used to estimate the probability of relevance for a given document, based on its retrieval status value and the natural logarithm of its rank. The final list was sorted according to these estimates. As mentioned in Section 2.6, our evaluation will be based on the leaving-one-out evaluation strategy producing an unbiased estimator of the real performance.

Some authors suggested multilingual IR systems that do not rely on a merging procedure that usually tends to degrade the overall average precision. In this vein, we should mention the document translation approach, one that unifies the collection language by translating all documents into a common language (Braschler and Schäuble 2001, McNamee and Mayfield 2002). As another method of eliminating the merging phase, Chen (2002) proposed that we build a new request composed of all possible translations and use this multilingual query to search in a single collection composed of all documents written in different languages. An evaluation of this scheme did not reveal any improved performance compared to one

based on merging multiple monolingual runs. A similar proposal was made by Martínez-Santiago et al. (2003). As an alternative, Chen (2002) suggested translating the retrieved set of documents into the request language and resubmitting the query to this smaller set of automatically translated documents (monolingual search). Such an MLIR strategy will clearly, on average, result in improved precision, but requires extensive computation and increases response time.

Finally, it is known that each collection cannot be expected to supply pertinent articles for each request. For example, the English corpus does not contain relevant documents for eight queries (see Table 1). Based on this fact, it might prove interesting to suggest a selection procedure that is able to determine whether a given collection can provide pertinent documents, based on the current request. However, such a selection approach is not usually proposed in the CLEF workshops and the recent work of Braschler et al. (2003) can be viewed as an exception. In this study, the authors were trying to estimate the number of documents to be extracted from each individual collection based on the overlap between an extended query and the top-ranked items provided by each corpus.

4.2. Learning curve in merging strategy

Considering the best MLIR systems over the last three CLEF campaigns, we see that the round-robin interleaving system provided the best performances in 2000 (Braschler and Schäuble 2001). During the same campaign, McNamee et al. (2001) described a run using this merging strategy that resulted in an improved retrieval effectiveness compared to a second run based on a normalized score merging scheme. Gey et al. (2001) also suggested using a raw-score merging strategy. During the second CLEF campaign, Savoy (2002a) obtained the best performance based on a normalized score (combRSV%) merging approach and this same author suggested another normalized score merging denoted by combRSVnorm (see Table 3). Finally, during the last CLEF evaluation campaign, Savoy (2003) suggested using logistic regression in order to predict the relevance probability for each document, depending on the collection from which this document was extracted, including its rank and score. As an alternative, we also suggested a biased round-robin approach which extracted not one document per collection per round but one document for the French, English and Italian corpus and two from the German and Spanish collections, see also (Braschler et al. 2003). Such a merging strategy exploits the fact that the German and Spanish corpora possess roughly twice as many articles as do the other collections (see Table 1), under the assumption that relevant documents are uniformly distributed across collections. This hypothesis is not really respected in our test collection. For example, the prior probability that a randomly chosen document extracted from the French collection is relevant is $27.66/87{,}191 = 0.0003172$ and the same probability for an article extracted from the German corpus is $38.76/225{,}371 = 0.000172$. Thus this biased round-robin approach must be viewed as an heuristic, taking into account the size differences of the merged collections.

In order to measure the learning curve obtained from resolving the merging problem, Tables 10 regroup the two most significant merging strategies for each CLEF workshop,

using the CLEF 2002 test collection as a common denominator. In these tables, the round-robin interleaving scheme was used as the baseline when evaluating merging approaches based on manually translated requests (Tables 10(a) or 10(b) when also considering Rocchio's pseudo-relevance feedback (Buckley et al. 1996)) or machine based translated queries (Tables 10(c) or 10(d) when including a blind query expansion phase). When using the blind query expansion, we fixed $\alpha = 0.75$, $\beta = 0.75$ and the system was allowed to add m terms extracted from the k best ranked documents from the original query. The exact values of these two parameters are depicted in Table 10(b) and 10(d), and these optimal parameter settings seem to be collection-dependant.

From these tables, one can see that the raw-score merging approach does not provide interesting retrieval effectiveness due to the presence of incomparable document scores across the collection, at least in our implementation. For this merging scheme, the difference in average precision resulting from the round-robin approach is always statistically significant. However, Gey et al. (2001) or Chen (2002, 2003) suggested an IR scheme where document scores are more comparable across the collections, resulting in better retrieval effectiveness compared to our raw-score merging scheme.

In the last column, we also indicated the percentage of improvement over the baseline. For example, when considering machine translated queries in Table 10(c), we can see an enhancement of 2.1% to 9.2% in 2001 merging schemes compared to the round-robin approach and an improvement of 4.5% to 11.9% in 2002. When comparing manually (Table 10(a)) and machine based translated queries (Table 10(c)), we usually found a difference in average precision of around -7.5% (e.g., round-robin strategy, 33.85 vs. 31.16 (-7.9%) or 37.28 vs. 34.60 (-7.2%) when using the query expansion technique).

More generally, we found that the round-robin approach can be viewed as a good first approximation. A simple normalization procedure (e.g., dividing by the maximum document score or combRSV%) presents retrieval performances similar to the round-robin merging scheme. A more sophisticated normalization based on the range of document scores (combRSVnorm) usually results in a statistically significant and improved average precision compared with the round-robin scheme. A similar conclusion can be drawn for the new version of the CORI model ("CORI new") (Callan 2000). As an alternative, we considered the biased round-robin strategy which usually reveals significantly improved retrieval effectiveness when compared to the round-robin scheme.

Finally, the logistic regression merging strategy clearly improved retrieval effectiveness when compared to all other merging procedures, and its performance was statistically better than the round-robin scheme. Moreover, when evaluating our logistic approach, we did not use the same set of queries to estimate the value of the underlying coefficients of the logistic model and to evaluate the merging strategy (retrospective evaluation). In the current evaluation, the training set included all requests except the current query (the leaving-one-out evaluation strategy) which produced an unbiased estimator of the real performance. As a second example of this approach robustness, the coefficients computed according to the CLEF 2001 test collection proved to be really successful in the CLEF 2002 test collection (Savoy 2003). This statistical method was also used in another context by Gey et al. (2001, 2002) and Chen (2002, 2003). These authors used the same coefficients for their CLEF evaluation campaigns, leading to the conclusion that this statistical method may be used in practice.

5. Conclusion

Convinced that isolated retrieval effectiveness evaluations are not very useful, we have carried out experiments on various search strategies that were applied using different languages. Based on our current evaluations and on the most effective IR systems suggested by the various CLEF evaluation campaigns, we have found that:

- effective monolingual IR systems can work with various European languages, on the basis of algorithms suggested for the English language (see Tables 2 and 5). In our case, we chose the Okapi model, but a probabilistic model based on logistic regression (Gey 2001, 2002, Chen 2002, 2003) may also result in adequate retrieval performances across the different languages;
- when working with agglutinative languages such as German, Dutch or Finnish, a combined IR model (see Table 4) or an indexing scheme using a decompounding scheme (Chen 2002, 2003) may provide better retrieval effectiveness than a word-based indexing procedure;
- in proposing effective bilingual IR systems, we knew that the different automatic translation resources tended to provide translations that led to great variability in retrieval performance (see Tables 7 and 8). Because of this phenomenon, automatic query (or document) translations should be based on a combined approach (see Table 9) or at least on an automatic selection procedure able to ascertain the most appropriate translation source(s) for a given request;
- for effective multilingual searches, it seems better and simpler to cross language barriers by applying query translation approaches (Savoy 2003, Chen 2002, 2003). To merge the results provided by each language into a single output list, a normalization based on maximum and minimum document scores (combRSVnorm) can perform well in terms of retrieval effectiveness (see Tables 10(c) and 10(d)). Moreover, when a learning sample is available, the logistic regression approach is able to generate the best retrieval performance;
- based on the last three CLEF evaluation campaigns, we can describe the learning curve by the following values. In monolingual IR systems, we improved average precision by 1.4% to 11.3%, depending on the language (see Table 5). For bilingual IR systems and depending on the language, retrieval effectiveness was increased by 7.8% to 18.2% in 2001 and 12.3% to 37.8% in 2002, as compared to the translation strategy used in 2000 (see Table 9). When proposing more effective MLIR systems, the enhancement was around 7% to 9.2% in 2001 and 11.9% to 15% in 2002, when compared to the CLEF 2000 merging approach (see Tables 10(c) and 10(d)).

Of course, these findings still need to be confirmed using other languages or other test collections. For the future, we need to improve our stopword lists and stemming procedures for European languages other than English and French. For those languages having high frequencies of compound word constructions, it is still worthwhile to know whether n-gram indexing approaches could achieve higher levels of retrieval performance than would enhanced word segmentation heuristics. An alternate question is

whether or not data fusion will remain the most effective search model for agglutinative languages.

In designing more effective bilingual IR systems, exploring the possibility of automatically selecting the most appropriate translation alternative seems to be worthwhile when faced with various translation resources. It would also seem advantageous to continue our investigations on statistical translation tools (Nie et al. 1999, Kraaij 2002) or a similarity thesauri (Braschler et al. 2002) that can be used to automatically translate the submitted query for less widely used languages (e.g., Swedish, Finnish) for which freely translated resources are not always available. Moreover, we could consider weighting some translation alternatives differently (Chen 2003) or at least conducting more evaluations on pre-translation query expansions (McNamee et al. 2001).

Finally, when searching in multiple collections that contain documents written in various languages, it is worthwhile to look at better collection merging strategies or to include intelligent selection procedures in order to avoid searching through a collection that does not contain any relevant documents.

Appendix 1: Weighting schemes

To assign an indexing weight w_{ij} that reflects the importance of each single-term T_j in a document D_i, we might use various approaches as shown in Table A.1 in which n indicates the number of documents in the collection, t the number of indexing terms, df_j the number of documents in which the term T_j appears, the document length (the number of indexing terms) of D_i is denoted by nt_i, and avdl, b, k_1, pivot and slope are constants. For the Okapi weighting scheme, K represents the ratio between the length of D_i measured by l_i (sum of tf_{ij}) and the collection mean noted by avdl.

Table A.1. Weighting schemes.

bnn	$w_{ij} = 1$	nnn	$w_{ij} = tf_{ij}$
ltn	$w_{ij} = (\ln(tf_{ij}) + 1) \cdot idf_j$	atn	$w_{ij} = idf_j \cdot [0.5 + 0.5 \cdot tf_{ij} / \max tf_{i.}]$
Okapi	$w_{ij} = \frac{((k_1+1)\cdot tf_{ij})}{(K+tf_{ij})}$	npn	$w_{ij} = tf_{ij} \cdot \ln\left[\frac{(n-df_j)}{df_j}\right]$
lnc	$w_{ij} = \frac{\ln(tf_{ij})+1}{\sqrt{\sum_{k=1}^{t}(\ln(tf_{ik})+1)^2}}$	ntc	$w_{ij} = \frac{tf_{ij}\cdot idf_j}{\sqrt{\sum_{k=1}^{t}(tf_{ik}\cdot idf_k)^2}}$
dtc	$w_{ij} = \frac{(\ln(\ln(tf_{ij})+1)+1)\cdot idf_j}{\sqrt{\sum_{k=1}^{t}((\ln(\ln(tf_{ik})+1)+1)\cdot idf_k)^2}}$		
ltc	$w_{ij} = \frac{(\ln(tf_{ij})+1)\cdot idf_j}{\sqrt{\sum_{k=1}^{t}((\ln(tf_{ik})+1)\cdot idf_k)^2}}$		
dtu	$w_{ij} = \frac{(1+\ln(1+\ln(tf_{ij})))\cdot idf_j}{(1-\text{slope})\cdot\text{pivot}+\text{slope}\cdot nt_i}$		
Lnu	$w_{ij} = \frac{(\ln(tf_{ij})+1/\ln(1_i/nt_i)+1)}{(1-\text{slope})\cdot\text{pivot}+\text{slope}\cdot nt_i}$		

Acknowledgments

This research was supported in part by the SNSF (Swiss National Science Foundation) under grants 21-58 813.99 and 21-66 742.01. The author would like to thank the three anonymous referees for their helpful suggestions and remarks.

References

Amati G, Carpineto C and Romano G (2003) Italian monolingual information retrieval with PROSIT. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Cross-Language Information Retrieval and Evaluation, Springer-Verlag, Berlin, 2003, pp. 257–264.

Ballesteros L and Croft WB (1998) Resolving ambiguity for cross-language retrieval. In: Croft WB, Moffat A, van Rijsbergen CJ, Wilkinson R and Zobel J, Eds. Proceedings of the 21st International Conference of the ACM-SIGIR 1998, The ACM Press, New York, pp. 64–71.

Brand R and Brünner M (2003) Océ at CLEF 2002. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Cross-Language Information Retrieval and Evaluation, Springer-Verlag, Berlin, 2003, pp. 59–65.

Braschler M and Schäuble P (2001) Experiments with the Eurospider retrieval system for CLEF 2000. In: Peters C, Ed. Cross-Language Information Retrieval and Evaluation, LNCS #2069, Springer-Verlag, Berlin, 2001, pp. 140–148.

Braschler M and Peters C (2002) CLEF methodology and metrics. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Cross-Language Information Retrieval and Evaluation, LNCS #2406, Springer-Verlag, Berlin, 2002. pp. 394–404.

Braschler M, Ripplinger B and Schäuble P (2002) Experiments with the Eurospider retrieval system for CLEF 2001. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Cross-Language Information Retrieval and Evaluation, LNCS #2406, Springer-Verlag, Berlin, 2002, pp. 102–117.

Braschler M, Göhring A and Schäuble P (2003) Eurospider at CLEF 2002. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Cross-Language Information Retrieval and Evaluation, Springer-Verlag, Berlin, 2003, pp. 164–174.

Buckley C, Singhal A, Mitra M and Salton G (1996) New retrieval approaches using SMART. In: Harman DK, Ed. Proceedings of TREC-4, NIST Publication #500-236, Gaithersburg, 1996, pp. 25–48.

Callan JP, Lu Z and Croft, WB (1995) Searching distributed collections with inference networks. In: Fox EA, Ingwersen P and Fidel R., Eds. Proceedings of the 18th International Conference of the ACM-SIGIR The ACM Press, New York, pp. 21–28.

Callan JP (2000) Distributed information retrieval. In: Croft WB, Ed. Advances in Information Retrieval, Kluwer, Boston, pp. 127–150.

Chen A (2002) Multilingual information retrieval using English and Chinese queries. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Cross-Language Information Retrieval and Evaluation, LNCS #2406, Springer-Verlag, Berlin, pp. 44–58.

Chen A (2003) Cross-language retrieval experiments at CLEF-2002. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Cross-Language Information Retrieval and Evaluation, Springer-Verlag, Berlin, 2003, pp. 28–48.

Dumais ST (1994) Latent semantic indexing (LSI) and TREC-2. In: Harman DK, Ed. Proceedings TREC-2, NIST Publication #500-215, Gaithersburg, pp. 105–115.

Efron B and Tibshirani RJ (1993) An Introduction to the Bootstrap. Chapman & Hall, New-York.

Figuerola CG, Gómez R and Zazo Rodríguez AF (2002) Spanish monolingual track: the impact of stemming on retrieval. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Cross-Language Information Retrieval and Evaluation, LNCS #2406, Springer-Verlag, Berlin, pp. 253–261.

Fox EA and Shaw JA (1994) Combination of multiple searches. In: Harman DK, Ed. Proceedings TREC-2, NIST Publication #500-215, Gaithersburg, pp. 243–249.

Gachot DA, Lange E and Yang J (1998) The SYSTRAN NLP browser: an application of machine translation technology. In: Grefenstette G, Ed. Cross-Language Information Retrieval, Kluwer, Boston, pp. 105–118.

Gey F, Jiang H, Petras V and Chen A (2001) Cross-language retrieval for the CLEF collections—comparing multiple methods of retrieval. In: Peters C, Ed. Cross-Language Information Retrieval and Evaluation, LNCS #2069, Springer-Verlag, Berlin, pp. 116–128.

Gey FC, Jiang H and Perelman N (2002) Working with Russian queries for the GIRT, bilingual and multilingual CLEF tasks. In: Peters C, Braschler M, Gonzalo J and Kluck M, eds. Cross-Language Information Retrieval and Evaluation, LNCS #2406, Springer-Verlag, Berlin, pp. 235–243.

Harter SP (1975) A probabilistic approach to automatic keyword indexing: Part I. On the distribution of specialty words in a technical literature. Journal of the American Society for Information Science, 26:197–206.

Hiemstra D, Kraaij W, Pohlmann R and Westerveld T (2001) Translation resources, merging strategies, and relevance feedback for cross-language information retrieval. In: Peters C, Ed. Cross-Language Information Retrieval and Evaluation, LNCS #2069, Springer-Verlag, Berlin, pp. 102–115.

Hosmer DW and Lemeshow S (2000) Applied Logistic Regression, 2nd Edn. John Wiley & Sons, New York.

Hull D (1993) Using statistical testing in the evaluation of retrieval experiments. In: Korfhage R, Rasmussen E and Willett P, Eds. Proceedings of the 16th International Conference of the ACM-SIGIR'93, The ACM Press, New York, pp. 329–338.

Kleinbaum DG and Klein M (2002) Logistic Regression, 2nd edn. Springer-Verlag, New York.

Kraaij W (2002) TNO at CLEF 2001: Comparing translation resources. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Cross-Language Information Retrieval and Evaluation, LNCS #2406, Springer-Verlag, Berlin, pp. 78–93.

Kwok KL, Grunfeld L and Lewis DD (1995). TREC-3 ad-hoc, routing retrieval and thresholding experiments using PIRCS. In: Harman DK, Ed. Proceedings TREC-3, NIST Publication #500-225, Gaithersburg, 1995, pp. 247–255.

Kwok KL, Grunfeld L, Dinstl N and Chan M (2001) TREC-9 cross-language, web and question-answering track experiments using PIRCS. In: Voorhees EM and Harman DK, Eds. Proceedings TREC-9. NIST Publication #500-249, Gaithersburg, pp. 417–426.

Le Calvé A and Savoy J (2000) Database merging strategy based on logistic regression. Information Processing & Management, 36:341–359.

Lovins JB (1968) Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 11:22–31.

MacFarlane A (2003) PLIERS and Snowball at CLEF 2002. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Cross-Language Information Retrieval and Evaluation. Springer-Verlag, Berlin, 2003, pp. 321–335.

McNamee P, Mayfield J and Piatko C (2001) A language-independent approach to European text retrieval. In: Peters C, Ed. Cross-Language Information Retrieval and Evaluation, LNCS #2069, Springer-Verlag, Berlin, pp. 129–139.

McNamee P and Mayfield J (2002) JHU/APL experiments at CLEF: translation resources and score normalization. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Cross-Language Information Retrieval and Evaluation, LNCS #2406, Springer-Verlag, Berlin, pp. 193–208.

Martínez-Santiago F, Martín M A and Ureña A (2003) SINAI at CLEF 2002: experiments with merging strategies. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Cross-Language Information Retrieval and Evaluation. Springer-Verlag, Berlin, 2003, pp. 187–196.

Molina-Salgado H, Moulinier I, Knudson M, Lund E and Sekhon K (2002) Thomson legal and regulatory at CLEF 2001: Monolingual and bilingual experiments. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Cross-Language Information Retrieval and Evaluation, LNCS #2406, Springer-Verlag, Berlin, pp. 226–234.

Monz C and de Rijke M (2002) Shallow morphological analysis in monolingual information retrieval for Dutch, German and Italian. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Cross-Language Information Retrieval and Evaluation, LNCS #2406, Springer-Verlag, Berlin, 2002, pp. 262–277.

Nie JY, Simard M, Isabelle P and Durand R (1999) Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In: Hearst M, Gey F and Tong R, Eds. Proceedings of the 22nd International Conference of the ACM-SIGIR 1999, The ACM Press, New York, pp. 74–81.

Nie JY, Simard M and Forster G (2001) Multilingual information retrieval based on parallel texts from the web. In: Peters C, Ed. Cross-Language Information Retrieval and Evaluation, LNCS #2069, Springer-Verlag, Berlin, pp. 188–201.

Nie J and Simard M (2002) Using statistical translation models for bilingual IR. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Cross-Language Information Retrieval and Evaluation, LNCS #2406, Springer-Verlag, Berlin, pp. 137–150.

Oard DW, Levow G-A and Cabezas CI (2001) CLEF experiments at Maryland: statistical stemming and backoff translation. In: Peters C, Ed. Cross-Language Information Retrieval and Evaluation, LNCS #2069, Springer-Verlag, Berlin, pp. 176–187.

Porter MF (1980) An algorithm for suffix stripping. Program, 14:130–137.

Powell AL, French JC, Callan J, Connell M and Viles CL (2000) The impact of database selection on distributed searching. In: Belkin NJ, Ingwersen P and Leong M-K, Eds. Proceedings of the 23rd International Conference of the ACM-SIGIR 2000, The ACM Press, New York, pp. 232–239.

Robertson SE and Sparck Jones K. (1976) Relevance weighting of search terms. Journal of the American Society for Information Science, 27:129–146.

Robertson SE and Walker S (1994) Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In: Croft WB and van Rijsbergen CJ, Eds. Proceedings of the 17th International Conference of the ACM-SIGIR'94, Springer-Verlag, London, pp. 232–241.

Robertson SE, Walker S and Beaulieu M (2000) Experimentation as a way of life: Okapi at TREC. Information Processing & Management, 36:95–108.

Salton G and McGill MJ (1983) Introduction to Modern Information Retrieval. McGraw-Hill, New York.

Salton G and Buckley C (1988) Term weighting approaches in automatic text retrieval. Information Processing & Management, 24:513–523.

Savoy J (1997) Statistical inference in retrieval effectiveness evaluation. Information Processing & Management, 33:495–512.

Savoy J (1999) A stemming procedure and stopword list for general French corpora. Journal of the American Society for Information Science, 50:944–952.

Savoy J and Rasolofo Y (2001) Report on the TREC-9 experiment: link-based retrieval and distributed collections. In: Voorhees EM and Harman DK, Eds. Proceedings TREC-9, NIST Publication #500-249, Gaithersburg, pp. 579–588.

Savoy J (2002a) Report on CLEF-2001 experiments: effective combined query-translation approach. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Cross-Language Information Retrieval and Evaluation, LNCS #2406, Springer-Verlag, Berlin, pp. 27–43.

Savoy J (2002b) Recherche d'informations dans des corpus en langue française : Utilisation du référentiel Amaryllis. TSI, Technique et Science Informatiques, 21:345–373.

Savoy J (2003) Report on CLEF-2002 experiments: combining multiple sources of evidence. In: Peters C, Braschler M, Gonzalo J and Kluck M, Eds. Cross-Language Information Retrieval and Evaluation. Springer-Verlag, Berlin, 2003, pp. 66–90.

Singhal A, Choi J, Hindle D, Lewis DD and Pereira F (1999) AT&T at TREC-7. In: Voorhees EM and Harman DK, Eds. Proceedings TREC-7, NIST Publication #500-242, Gaithersburg, pp. 239–251.

Sproat R (1992) Morphology and Computation. The MIT Press, Cambridge.

van Rijsbergen CJ (1979) Information Retrieval, 2nd edn. Butterworths, London.

Voorhees EM, Gupta NK and Johnson-Laird B (1995) The collection fusion problem. In: Harman DK, Ed. Proceedings TREC-3, NIST Publication #500-225, Gaithersburg, pp. 95–104.