Information Retrieval, 7, 183–204, 2004

Combination Approaches for Multilingual Text Retrieval

MARTIN BRASCHLER martin.braschler@eurospider.com
Eurospider Information Technology AG, Schaffhauserstrasse 18, CH-8006 Zurich, Switzerland; Université de Neuchâtel, Institut Interfacultaire d'Informatique, Pierre-à-Mazel 7, CH-2001 Neuchâtel, Switzerland

Received December 4, 2002; Revised May 14, 2003; Accepted May 14, 2003

Abstract. We describe the Eurospider component for Cross-Language Information Retrieval (CLIR) that has been employed for experiments at all three CLEF campaigns to date. The central aspect of our efforts is the use of combination approaches, effectively combining multiple language pairs, translation resources and translation methods into one multilingual retrieval system. We discuss the implications of building a system that allows flexible combination, give details of the various translation resources and methods, and investigate the impact of merging intermediate results generated by the individual steps. An analysis of the resulting combination system is given which also takes into account additional requirements when deploying the system as a component in an operational, commercial setting.

Keywords: Cross-Language Information Retrieval, combination methods, merging, operational systems

1. Introduction

The Eurospider components for Cross-Language Information Retrieval (CLIR) rely heavily on the combination of multiple simpler CLIR approaches. We discuss the implications of combining different translation methods, and present the supporting framework that we have implemented. Eurospider has participated with considerable success in all three CLEF campaigns held to date, and we analyze the results that have been obtained.

Definitions

The CLEF campaign offers both multilingual and bilingual cross-language information retrieval (CLIR) evaluation. Multilingual retrieval in CLEF is defined as the task of retrieving documents from a multilingual collection containing articles from newspapers, newsmagazines and news agencies written in any of four (English, French, German, Italian in CLEF 2000) or five (as before, plus Spanish in CLEF 2001 and CLEF 2002) languages. Bilingual retrieval is a "simpler" task, retrieving documents from a monolingual collection written in a language different from the one used for query formulation.

Overview

While initial experiments are much older (see e.g. Salton 1970), interest in CLIR has substantially increased after 1996. The CLIR approaches that have subsequently been proposed

can be classified in different ways. Braschler et al. (1998a) suggested a classification in terms of what information is translated to "bridge the language gap" in CLIR: the queries (query translation; QT), the documents (document translation; DT), or both.[1] Oard (1997) divided approaches into those using controlled vocabulary vs. those working on free text, then further refining this classification by the type of resources used for classification: corpus-based or knowledge-based.

Current state-of-the-art translation approaches to free-text CLIR, such as those used by the vast majority of CLEF participants, fall broadly speaking into three categories: approaches using machine translation (MT) systems, approaches using machine-readable dictionaries (MRD) and approaches using corpus-based resources derived from suitable training data. Eurospider has used methods from all three fields in the TREC and CLEF campaigns.

Currently, query translation (QT) seems to be more popular than document translation (DT), mainly because of perceived limitations in scalability in the latter approach. We have attempted the combination of both methods.

Combination of Multiple Simpler Approaches

Much of the early work on CLIR after 1996 obtained an effectiveness that was far below what was observed for monolingual retrieval (for an overview of this type of work, see e.g. Grefenstette 1998). This substantial drop-off in effectiveness was found to be fairly consistent across different approaches, such as using MT or MRDs. As a consequence, we started early on to search for combination translation approaches in the hope of being able to substantially narrow the gap between CLIR and monolingual retrieval (Braschler and Schäuble 1998, Braschler et al. 1998b). Using a combination approach is analogous to efforts in monolingual IR to combine multiple sources of evidence (Belkin et al. 1993b, Fox and Shaw 1993).

The combination of multiple translation methods also decreases the risk of retrieval failure due to missing or incorrect translations. A substantial part of the difference in average retrieval effectiveness between monolingual and cross-language IR can be attributed to a few (negative) outliers. Such outliers are less likely if multiple sources of translations are used, and retrieval performance becomes more robust. A combination of multiple translation resources makes it possible to greatly increase the lexical coverage of a CLIR system, which is essential if a wide range of queries are to be processed, including technical terminology, fashionable terms and names.

Merging of Different Languages

When going from bilingual CLIR to truly multilingual CLIR, involving more than one target language, the additional question arises of whether to handle the multiple target languages simultaneously or each language separately and then later merge the individual results. Systems that handle all languages simultaneously often have to translate all items into some form of common interlingua in order to build an unified index. By handling the languages in pairs this added difficulty can be avoided at the expense of having to deal with a number of intermediate results. However, combination approaches lend themselves naturally to this

type of architecture, as they typically already require merging of intermediate results coming out of the different types of translation resources.

In both cases, serious issues arise with respect to weighting between the different languages, an aspect that is as such absent in the monolingual case, although there are parallels to the monolingual problem of merging databases indexed by different, autonomous systems (see e.g. Du and Callan 1998). If attempting to merge results from several bilingual retrieval runs on documents in different languages, the estimates of relevance obtained through the weighting scheme will typically not be comparable, since the different vocabularies of the languages yield different term frequencies. Similarly, when using a single, unified index to permit parallel retrieval, different numbers of documents across the languages will produce incompatible frequency counts.

Key Requirements

In summary, a combination approach to multilingual CLIR in CLEF has to address

1. how to combine different translation approaches, and which types of resources to use (MRD, MT, corpus-based)
2. what to translate (query, documents, both, none)
3. how to handle the multiple languages (simultaneously or separately)

Related work

Overviews on different approaches to cross-language information retrieval can be found e.g. in Braschler et al. (1998) and Oard (1997). Good examples of the use of all three main types of translation resources commonly used for CLIR (machine translation, machine-readable dictionaries and corpus-based data structures) can be found in experiments by CLEF participants. For machine translation, see e.g. Jones and Lam-Adesina (2002) and Figuerola et al. (2001), for machine-readable dictionaries and thesauri, see e.g. Hedlund et al. (2001) and Gey et al. (2002), and for corpus-based CLIR see e.g. Hiemstra et al. (2001) and Nie and Simard (2002).

Combination approaches, such as those used in our system, make use of two or more of these types of translation resources. University of California at Berkeley has also used combination approaches since CLEF 2000 with considerable success (Gey et al. 2001, Chen 2002a). Further experiments with combination approaches were also carried out by Université de Neuchâtel (Savoy 2002a, 2002b) and Thomson Legal (Moulinier and Molina-Salgado 2002). All these papers also address the question of how to merge the intermediate results obtained by the individual components of the respective systems. Today, a majority of participants in the multilingual track of CLEF use combination approaches, including the groups scoring in the top three in terms of effectiveness. Our work differs from the work of most other CLEF participants in that we have used translation resources of all three types: MT, MRD and corpus-based, and that we have combined both query translation and document translation, resulting in a broad understanding of the properties of such combinations and in a system and a framework that can handle the widest possible range of translation resources. While Chen (2002a) reports on the combined use of bilingual dictionaries, parallel

corpora and mining of search engine results for query translation, he does not use document translation. Other CLEF participants to try a combination of document and query translation are McNamee and Mayfield (2002b), but this was in unofficial experiments. They used a bilingual dictionary derived from a parallel corpus for translation.

Furthermore, our system is built with applications in operational, commercial settings in mind, therefore emphasizing some issues that we identified as being important in such situations, such as good lexical coverage, the avoidance of negative outliers and applicability to a broad range of domains. These criteria are usually not in the main focus of the CLEF evaluation.

In monolingual information retrieval, there is also earlier work which has relevance to our experiments on combination of approaches and on merging of languages. In these studies, the corresponding problems are also referred to as "data fusion" (merging of multiple result lists calculated on the same document collection), "collection fusion" (merging of multiple result lists calculated on disjoint document collections), and "query combination" (merging of multiple query representations before retrieval). Much of the work on combining multiple result lists coming from the same document collection stems from the observation that the estimation of relevance of documents improves when multiple sources of evidence are available. Saracevic and Kantor (1988) observed in their studies with multiple manually generated query formulations that "[..] the more often an item was retrieved the more the odds shifted in favor of relevance". Similar work by Belkin et al. (1993a) with multiple Boolean query formulations showed that "progressive combination of query formulations leads to progressively improving retrieval performance". The same report also gives a good overview of early research on this phenomenon (see also Belkin et al. 1993b, Bartell et al. 1994). Early experiments on combining the output of different weighting schemes in different ways were conducted by Fox et al. (1992), Fox and Shaw (1993) and Lee (1995), generally with encouraging results. The problem of merging result lists from disjoint document collections, namely those formed by the different target languages, is closely related to the problem of merging output from multiple, separate document collections in the monolingual case. Earlier work on the monolingual problem can be found e.g. in Callan et al. (1995) and Voorhees et al. (1995), both of which also provide good introductions to the problem.

2. Individual translation resources and approaches

2.1. Machine translation (MT)

Machine translation systems, i.e. systems that attempt to automatically translate texts from one language to another, seem like an obvious choice for CLIR. However, there are several obstacles to using MT for CLIR. Firstly, MT systems attempt to determine the correct word sense for translation by using context analysis. If an MT system is applied to query translation, there may be little or no context available that the can be exploited. Furthermore, MT systems depend on correct, grammatical input. Queries are often only a string of search terms enumerated by the user. Lastly, the diversity of concepts entered by a user is nearly endless—users search for names, current affairs, rare items, "fashionable" terminology,

etc.—all things that are hard to translate for an MT system that usually uses some form of dictionary as resource. MT would appear to be a more attractive choice for document translation—there is more context available to be exploited in a document, and grammar is typically cleaner. However, the main problem with document translation remains—if the MT system cannot translate a certain term, later searches will come up without result.

Eurospider has used MT systems both for document translation (DT) and query translation (QT) in CLEF experiments. For DT, all documents were processed offline, and then inserted in their translated form into the system. QT would require a direct integration of the MT and retrieval systems. We have opted for a looser coupling by externally translating the query and then feeding the result to the retrieval system. No manual modifications to the query were made of any kind.

MT systems are only available for a limited number of "popular" language pairs. For our CLEF experiments, we were not able to locate a suitable German/Spanish MT system, so German/Spanish MT was simulated by first using MT to translate to English, and then having the system translate the resulting English output to Spanish. We feel this illustrates well the kind of difficulties that will be faced when attempting to deploy a CLIR system for lesser used language pairs. Clearly, MT has deficits in such situations.

2.2. Similarity thesaurus (ST)

We have used corpus-based approaches with considerable success for CLIR. Corpus-based resources are automatically derived from suitable training data. Our experiments with corpus-based methods in CLEF use a so-called "similarity thesaurus" (ST) (Qiu and Frei 1993).

To build a multilingual similarity thesaurus, the training data is constructed to consist of pairs of documents with related content in the desired languages. For example, one pair of documents may consist of one document each for both languages about election results (Braschler and Schäuble 1998). Such document pairs can be found by obtaining collections of similar type, e.g. news articles in several languages, and then "aligning" individual documents. The alignment process works by picking out "indicators", i.e. special characteristics of a document, and comparing them with the indicators of documents in the other language. Indicators can be of different type, e.g.

1. Date, author, source of a document
2. Classifiers assigned to a document, if available
3. Proper names in documents that are spelled (nearly) identically across languages
4. Numbers and acronyms
5. Words contained in a small core vocabulary for which a dictionary is available before the alignment process ("seed dictionary").

We have shown that it is possible to align documents from independent sources based only on indicators such as those described above (see figure 1) as long as the documents

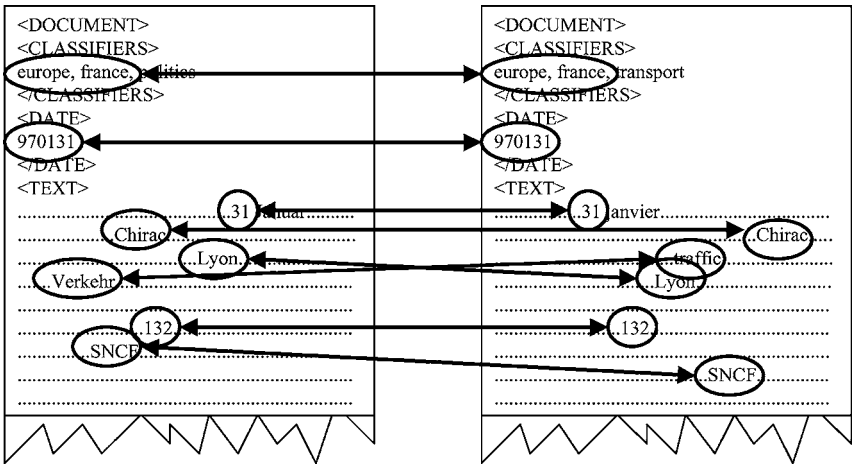

Figure 1. Alignment of documents. The two documents are aligned based on matching classifiers, dates, numbers, names, and a few terms coming from a seed dictionary.

are from the same domain (Braschler and Schäuble 1998). The similarity thesaurus construction process that relies on the training data produced through this form of document alignment is relatively robust with regard to the quality of the alignments. There is no need for alignment on paragraph or even sentence level. It is sufficient if documents that are paired treat loosely the same subject. This is an essential advantage compared to some other corpus-based approaches that rely on training on so-called "parallel corpora", which contain every item in all languages that are to be covered. While it is hard and usually expensive to obtain this form of parallel training collection, it is substantially easier to find related collections in the necessary languages, and convert them to training data through the alignment process. Interesting developments that may help increase the availability of suitable parallel corpora for training comes from work on mining the World Wide Web for suitable documents (Hiemstra et al. 2001, Nie and Simard 2002). We will explore training the similarity thesaurus on such corpora in the future.

The similarity thesaurus is constructed by calculating the similarity of every term in one language with all terms in the other language based on their occurrences in pairs of aligned documents. This process is similar to classical document retrieval. Where for document retrieval terms are used to find the documents they are contained in, we find those terms that are represented by the same documents, essentially using the documents to retrieve the terms. Indeed, it is possible to use the concept of a "dual space", where the roles of documents and terms are reversed for retrieval, and then use classical IR weighting schemes in their adapted form for similarity calculation. We use the standard *tf.idf* weighting adapted for the dual space ("*ff.iif*") to calculate the similarities between terms (see Formula 1). For every term, the top thirty most similar terms in the target language are then saved to the resulting thesaurus.

$$sim(\varphi_h, \varphi_i) := \frac{\sum_{d_j \in \varphi_h \cap \varphi_i} a_{j,h} * a_{j,i}}{\sqrt{\sum_{d_j \in \varphi_h} \left(a_{j,h}^2\right)} * \sqrt{\sum_{d_j \in \varphi_i} \left(a_{j,i}^2\right)}},$$

$$a_{j,i} := ff(d_j, \varphi_i) * iif(d_j), \quad iif(d_j) := \log\left(\frac{1+\Phi}{1+|d_j|}\right)$$

Formula 1. The *ff.iif* weighting used for similarity thesaurus calculation. The similarity between two terms is the angle between two vectors composed of *ff.iif* weights, i.e. the product of the feature frequency and of a value that is analogous to the *idf* used in classical information retrieval.

Since these "similarities" are derived on a purely statistical basis, there is no guarantee that actual translations are found. Typically, however, terms are found that describe a meaning identical or very closely related to that of the original term, but in the target language. Clearly, such terms can then be used to search for the concept expressed by the query. An added benefit of the statistical approach underlying similarity thesaurus construction is that it is not only possible to calculate similarities on a term/term basis, but also for multiple terms in one go—capturing the "concept" of the entire query as a whole. In this way, some of the ambiguities of individual terms can be avoided when the query contains more than one search term.

Probably the most exciting aspect of corpus-based resources is that when deriving suitably powerful resources, such as the similarity thesaurus, the coverage is only limited by the scope of the training data. By feeding large amounts of appropriate training data, the resource can cover vocabulary that goes far beyond the core vocabulary usually covered by even the largest manually constructed translation resources. Even more interesting, this extra vocabulary can contain items such as names, technical terminology and "fashionable expressions", which are usually very important for operational search systems.

2.3. Machine-readable dictionaries (MRD)

Machine-readable dictionaries (MRDs) are close relatives to the usual, well-known printed bilingual dictionaries. They contain for each entry in the source language one or more possible translations. Before use in retrieval systems, extra annotations present in printed dictionaries, such as case or gender information, must be removed, and abbreviations have to be expanded.

The main problem in using machine-readable dictionaries for CLIR is the ambiguity of many search terms. A CLIR system using MRDs can either use all possible translations, relying on a robust weighting mechanism during subsequent retrieval, or attempt to chose the most suitable translation, either by using frequency information if available, or by employing word-sense disambiguation. Results regarding the performance of these alternatives are inconclusive, and seem to heavily depend on the subsequent retrieval process.

We have used MRDs only in CLEF 2001, where we found no additional benefit with respect to using a combination of machine translation and similarity thesauri. This is probably due to the fact that an MT system tends to provide a similar lexical coverage to that of a typical MRD covering general terminology. We will not investigate MRDs in-depth in this paper.

However, the integration of MRDs is essential for operational settings, where customer-specific translation resources in the form of dictionaries and thesauri have to be employed.

3. Combination approaches

The Eurospider CLIR components allow flexible combination of multiple translation resources and retrieval approaches. By allowing such combinations, we aim to

1. *Maximize lexical coverage for translation*. Lexical coverage remains a central problem in CLIR systems, since a query that cannot be translated results in what is essentially worthless retrieval results (if retrieval results are returned at all). Not only will average performance of the system suffer in CLEF-style evaluations, but the acceptance of real-world users for systems that cannot handle a substantial portion of queries is low—users are very sensitive to the worst experiences they have with a system (especially those made early in their use of the system). Furthermore, a free-text retrieval system is typically faced with a wide range of different search requests, from well-formed grammatical sentences to a few hastily entered terms mixed from names and popular expressions. Not all forms of translation resources are equally suited to handle these different requests.
2. *Use multiple translation approaches/resources to avoid negative outliers*. Lexical coverage is central to avoiding negative outliers. However, such outliers can also stem from inappropriate translation. Translating the query using multiple different translation resources and approaches minimizes the chance of very pronounced outliers. There is a danger, however, of also minimizing the effect of positive outliers.
3. *Maximize the use of all available resources, allowing broad applicability to a range of operational settings*. Beyond some of the most popular languages pairs, usually involving English, resources are still typically scarce. A system should be able to work with whatever is available. Similarly, there are few resources covering special terminology and fashionable terms. Again, a system should be ready to use whatever is available.

The last point is especially challenging, in that it requires a system that can potentially work with only a fraction of the resources that are available for popular language pairs.

The Eurospider system allows flexible combination of translation resources and approaches either through direct combination of translation resources, or through the merging of intermediate retrieval results.

Combination approaches have shown to be effective for multilingual retrieval in the past CLEF campaigns, with the best entries for the 2002 multilingual track coming from groups using such approaches (Savoy 2002b, Chen 2002b, Braschler et al. 2002b).

3.1. Direct combination of translation resources

Two or more translation resources are directly combined into one "meta-resource" that comprises the union of the information contained in the individual resources. To make a combination in this sense possible, the resources must be "compatible", i.e. they should use matching types of input and output. Secondly, the resources must be available in a format that allows combination.

We have so far experimented with the direct combination of multiple machine-readable dictionaries, of multiple similarity thesauri, and a combination of machine-readable dictionaries and similarity thesauri. Only the combined MRDs were used in our CLEF experiments.

The direct combination is attractive, because it eliminates duplicate information, and therefore duplicate translation steps, and it avoids the need for later merging of intermediate results. When directly combining dissimilar translation resources, such as similarity thesauri and MRDs, weighting of the individual parts is a problem. This weighting problem is not dissimilar from the problem incurred when using the translation resources separately, and merging their individual outputs—in both cases the weighting should reflect the potential of the corresponding translation resources to contribute to a good overall result. We therefore avoid the direct combination of incompatible translation resources, using the merging of their individual outputs instead.

In order to allow direct combination of resources, Eurospider has developed a framework that operates on a common XML format. We use ISO Latin-1 for encoding. A good framework for direct combination is especially helpful in operational settings, where customer-specific translation resources are available and must be integrated into the system's larger, general-purpose resources.

3.2. Combination of intermediate results

Alternatively to building one large multilingual search index, a multilingual text retrieval system may handle languages in pairs, where the intermediate results from each language pair are later merged into one unified result. This is the approach chosen by most participants in the CLEF multilingual track.

In an analogous way, it is also possible to use the different translation resources individually, and then later merge the resulting intermediate results into one combined result that contains the results obtained by using all of the resources. The resources proper need not be combined directly in this approach, making it possible to use resources of very different types that are not compatible for direct combination.

Eurospider has combined outputs from machine translation, from similarity thesaurus translation and from MRDs in this way. The merging approaches used for such combinations are the same as for the combination of different language pairs , and are detailed in Section 4.

3.3. Architecture of a combination approach

The Eurospider CLIR component allows the following combinations:

– Multiple language pairs
– Document translation and query translation
– Multiple query translation approaches and resources

The systems we used for CLEF experiments work by first running each individual query translation method for each individual language pair. At this stage, the system mirrors a collection of classical bilingual CLIR systems using different translation resources. It is

therefore possible to use all refinements developed for such bilingual CLIR, an important advantage of this form of combination approach. Specifically, we use post-translation blind relevance feedback, which was shown to be effective in Ballesteros and Croft (1996). Based on our experience in all CLEF campaigns, we currently use the 10 best terms from the 10 top-ranked documents and Smart ltc weights for expansion. For the actual retrieval, we use the Smart Lnu.ltn weighting scheme, which has repeatedly been shown as effective for this type of document collections (Singhal et al. 1996).

When the intermediate results for all language pair/query translation method combinations are available, the first merging step is executed. We have experimented with two options at this stage: merging all outputs of one translation method into a multilingual result (see figure 2), or merging all outputs for one target language into a monolingual result combining the different translation approaches (see figure 3).

In a second operation, the system then either merges the multilingual intermediate results into one multilingual result that covers all translation methods, or it merges the different monolingual results into a multilingual result. In both cases, the result covers all query translation methods and all language pairs.

We have found only small performance differences in our experiments with both alternatives, with first merging all languages, and then the resources in a second step being slightly beneficial for CLEF 2002.

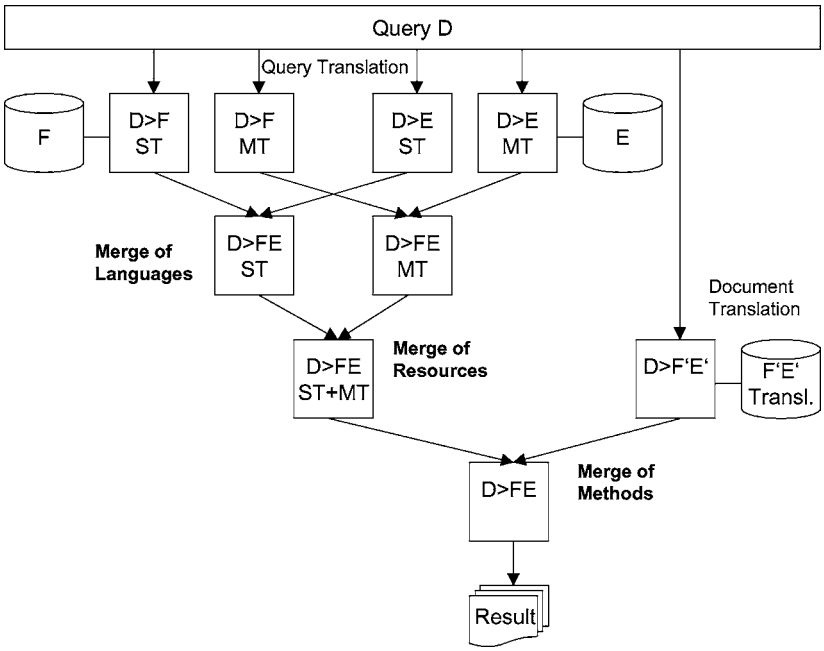

Figure 2. Combination system for CLIR, in this example for a system translating from German (D) to French (F) and English (E), combining different language pairs (merge of languages), different translation resources (merge of resources, machine translation "MT" and similarity thesauri "ST") and lastly different translation methods (merge of methods, query translation "QT" and document translation "DT").

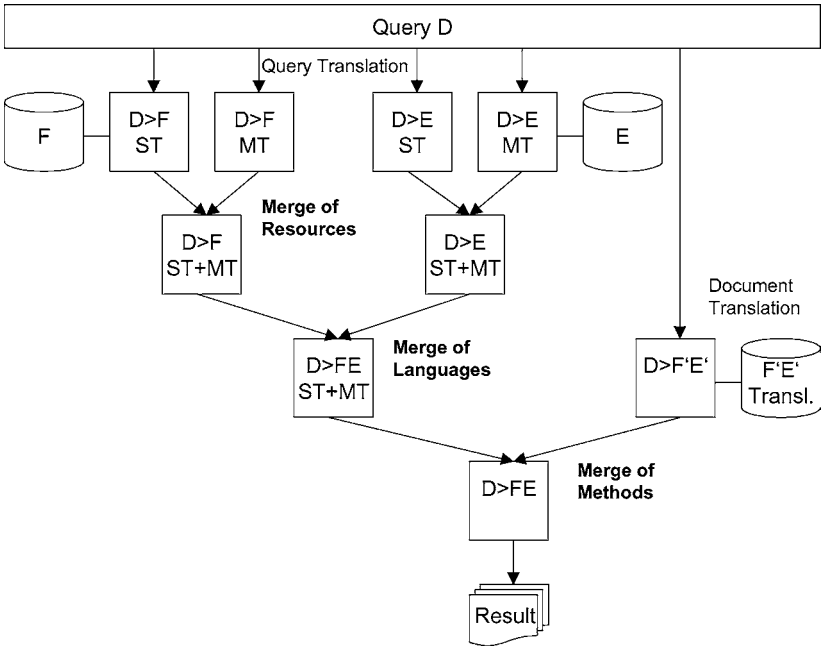

Figure 3. Combination system for CLIR, in this example for a system translating from German (D) to French (F) and English (E), combining different translation resources (merge of resources, machine translation "MT" and similarity thesauri "ST"), different language pairs (merge of languages) and lastly different translation methods (merge of methods, query translation "QT" and document translation "DT").

In a final merging step, the result from query translation is then merged with a result obtained from document translation. The document translation result is produced by inserting all translated documents into the system and then doing simply monolingual retrieval on this index.

4. Merging

Clearly, merging is an integral part in our combination approach to CLIR. In earlier CLEF participations, we have used two simple merging strategies: rank-based merging and interleaving.

For merging, we essentially distinguish two scenarios:

1. Both retrieval results were calculated on the same search space. The two result lists will essentially be a reordering of each other (with some extra items appearing at the bottom of the lists). This is the case if the output from experiments using the same language pair and different translation resources is merged (analogous to "data fusion" in monolingual retrieval, see e.g. Belkin et al. 1993a, Fox and Shaw, 1993).

2. The sets of documents in the result lists are disjoint. This is the case if the runs were produced through retrieval on disjoint search spaces, e.g. one search on the English part of the multilingual CLEF collection, and another search on the French part (analogous to "collection fusion" in monolingual retrieval, see e.g. Callan et al. 1995, Voorhees et al. 1995).

Some merging strategies apply to both scenarios, whereas some strategies can only be used for the first scenario. Rank-based merging can only be applied if the search spaces are shared among the lists to be merged. Interleaving is a more general strategy and can be used for both scenarios.

The main difficulty in merging is the lack of comparability of scores across different result lists. Result lists obtained from different collections, or through different weightings, are commonly not directly comparable. The retrieval status value RSV that the weighting scheme attaches to every document is only used for sorting the list, and is only valid in the context of the query, weighting and collection used.

4.1. Simple merging strategies

The two merging strategies described below address the problem of incompatible RSVs by not using the original RSV scores at all in determining the new rank of a document in the merged result list.

Rank-based Merging

For rank-based merging, calculation of a new RSV value for the merged list is based on the ranks of the documents in the original result lists. To calculate the new RSV of a document, its ranks in all the result lists are added. There are variations in this way of doing rank-based merging, such as the proposal by Belkin et al. (1993b) to use the median of all the ranks that a document obtained in the various result lists.

Clearly, the strategy only applies if the search space of all runs is shared, and therefore a substantial "overlap" in the documents retrieved for the individual runs exists. Since we feel that there is more importance in a rank difference between highly ranked documents than in a similar difference among lower ranked documents, we introduced a logarithmic dampening of the rank value, thus boosting the influence of highly ranked documents.

As mentioned previously (Section 3), we want to minimize the number of negative outliers. By using the dampening function, a document that receives one very good and one very poor rank will be ranked higher in the merged result list than a document receiving two mediocre ranks. Good documents are therefore not excessively penalized if one of the two retrieval runs performs very poorly. Of course, this effect has two sides: on the other hand, highly ranked documents from the poor run are also not downweighted drastically. In consequence, positive outliers are also blunted.

Interleaving

As an alternative that applies in both merging scenarios (same search space and disjoint search space), we have in the past used interleaving (also known as "round robin"): the merged result list is produced by taking one document in turn from all individual lists. If the collections are not disjoint, duplicates will have to be eliminated after merging, for example by keeping the most highly ranked instance of a document. Interleaving is a fairly old strategy, and has frequently been used as a baseline for comparison with newer schemes before (see e.g. Callan et al. 1995, Voorhees et al. 1995).

4.2. Improved merging strategies

We started to investigate improved merging strategies for the CLEF 2002 campaign. Two new methods were investigated. The first is a slight update of the interleaving strategy. The second is more elaborate, and presents an attempt to guess how well a query "hit" a language-specific subcollection.

Collection Size Based Interleaving

One main deficiency of interleaving as described above is that all result lists are handled equally, taking the same number of documents from each. It is extremely difficult to determine the number of relevant documents to be expected in the individual subcollections, but we have observed large collections in CLEF contribute on average more relevant documents per query than the smaller ones. Consequently, we have used a simple update to the straight interleaving method: we set the portion of documents taken from any one result list to be proportional to the size of the corresponding subcollection.

Feedback Merging

The second new strategy aims to predict the amount of relevant information contributed by each subcollection for a specific search request. It does this by carrying out an initial retrieval step, and then analyzing the top ranked documents from the result set, building an "ideal" query to retrieve that set of documents. This query is then compared to the original query, determining the overlap as an indication of the degree to which the concepts of the original query are represented in the retrieval result. The better such representation, the higher is the estimate of relevant documents (see figure 4). The result lists are then finally merged in proportion to these estimates. The biggest advantage of this method is its query dependence: whereas all the other methods described above use fixed ratios for merging the different result sets, this method determines an "optimal" ratio per query.

5. Analysis

Eurospider has participated with combination systems in all three CLEF campaigns to date. The results in all cases compare favorably to other participants, with the systems

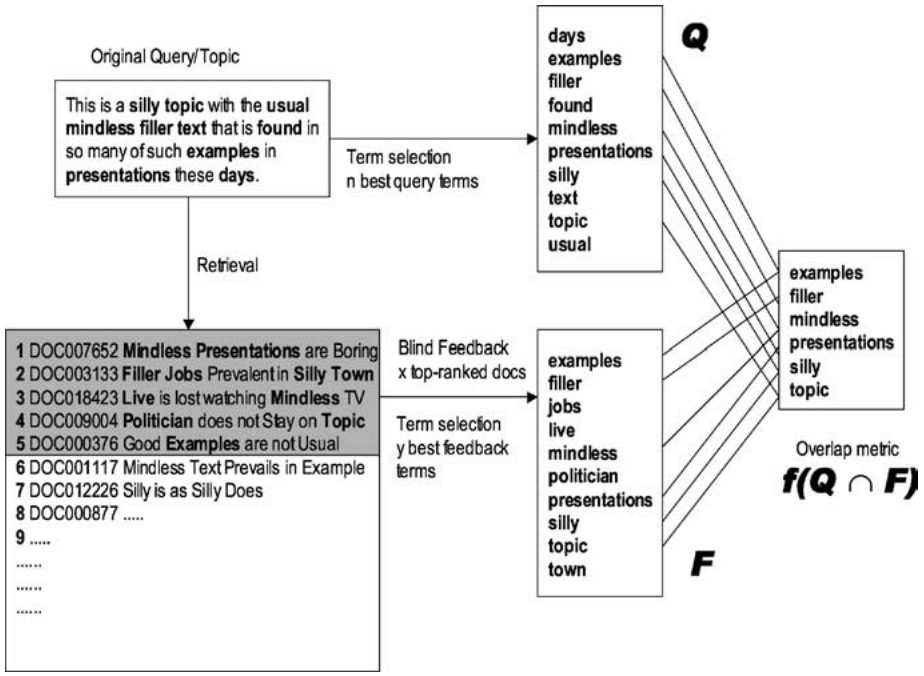

Figure 4. Simplified diagram of the feedback merging method. The query is used to obtain a list of highly ranked documents, from which representative terms are selected. High overlap of these terms with query terms indicates a high density of relevant items in the retrieval result.

placing among the top groups in all cases, and delivering the top performance for CLEF 2000.

Since the components are developed with integration into Eurospider's commercial retrieval products in mind, there are some design decisions that may contradict the objective of getting optimal performance on the CLEF corpus. In particular, we did not want to tune the system specifically with regard to the CLEF test collections in the following aspects:

1. Training of translation resources on the corpus itself. It was considered to be unrealistic for most operational cases that the collection itself can be used as training data.
2. Asymmetric handling of language pairs. There is a potential benefit in fine-tuning the system for every language pair. However, it is often unclear whether such tuning actually addresses peculiarities of a language pair or rather those of the underlying training collection. Again, it was considered unrealistic that extensive corpus-specific language tuning would be possible in operational settings.
3. Training on a previous year's relevance assessments. It is possible (and within the scope of CLEF permitted) to use the previous year's relevance assessments for fine-tuning of the system. However, by doing so, a situation is simulated where a lot of evaluation data is available for a corpus. We consider it to be unrealistic that corpus-specific evaluation

data of characteristics similar to a set of CLEF relevance assessments would be available in most operational settings.

All these three optimizations were reported to be potentially beneficial for evaluation at the last CLEF workshops (Rogati and Yang 2002, Brand and Brünner 2002, Savoy 2002b).

By taking the former three considerations into account, it follows that all our CLEF experimental systems were handling languages symmetrically to the extent possible, translation resources were trained on similar but disjoint training data, and relevance assessments from the previous years were not taken into account. We think the good performance of our systems in CLEF is all the more remarkable under these circumstances.

Besides the objective inherent to the CLEF evaluation of getting good performance with respect to the measures calculated from the relevance assessments, such as precision and recall, we were also interested in additional criteria which we consider essential for success in operational settings, such as:

1. Lexical coverage
2. Robustness with regard to negative outliers
3. Broad applicability of the system to a range of operational settings

By respecting these criteria, it may become necessary to add components to the combined approach that do not positively influence recall/precision measures or potentially even have a slightly negative impact.

In the next sections, we will comment on these three criteria with regard to the results observed at the three CLEF campaigns. Analysis of the individual results with respect to a single year's CLEF evaluation and their comparative performance with those of that year's participants is not the focus of this paper, and is treated in detail in previously published work (Braschler and Schäuble 2001, Braschler et al. 2002a, 2002b).

5.1. Lexical coverage

Typically, for CLEF topics, modern MT systems work well. This is especially true for long queries, which are essentially complete sentences, allowing MT systems to make use of their facilities for word sense disambiguation. The situation is a bit trickier for short queries, but these also work reasonably well as long as the terms come either from general vocabulary, or contain names that are not mistakenly translated. In the remaining cases, a corpus-based approach that, if appropriately trained, covers domain-specific terminology and fashionable terms, can help.

We have boosted lexical coverage of our experiments by using the corpus-based similarity thesaurus in conjunction with machine translation. As training data, we used additional news wire data that did not overlap with the newspaper and news wire data in the CLEF collections. Clearly, to be able to beneficially use a similarity thesaurus for our combination system, the training data must be sufficiently comparable in terms of terminology to the target document collection. In the optimal case, we could use the target collection for training data. As we mentioned above (Section 5), we think this is normally not possible in operational settings.

Table 1. Comparison of experiments using machine translation (MT), similarity thesauri (ST) and a combination of both. Shown are average precision figures. We only observed a clear benefit from combination for CLEF 2000.

	CLEF 2000	CLEF 2001	CLEF 2002
MT	0.2557	0.2929	0.3220
ST	0.1656	0.1778	0.1689
MT + ST	0.2622	0.2903	0.2876

Table 2. Direct per-query comparison of machine translation (MT), similarity thesauri (ST) and a combination of both.

	CLEF 2000	CLEF 2001	CLEF 2002
MT + ST vs. MT	23:17	19:31	18:32
MT + ST vs. ST	39:1	49:1	49:1
MT vs. ST	36:3	43:7	45:5

We consider the requirement of obtaining additional training documents from the same domain a much smaller obstacle and much easier to satisfy in operational settings. When looking at the results presented in Tables 1 and 2, a mixed picture emerges. Only in CLEF 2000 have we seen a benefit in terms of average precision when merging the output from the similarity thesaurus. However, this is mainly due to inconsistent quality of the thesauri we have used for the individual language pairs. We have not had access to suitable training data for English and Spanish, which led to similarity thesauri that did not perform as well as desired. A detailed analysis we performed for CLEF 2000 (Braschler and Schäuble 2001) shows that additionally to providing broader lexical coverage, a combination of machine translation and similarity thesauri can also boost average precision, when good-quality thesauri are available, as was the case for German/French and German/Italian.

5.2. Robustness with respect to negative outliers

Measures such as average precision, while very valuable for objective evaluation of systems in a setting such as CLEF, emphasize a goal of tuning the system for optimal performance on an average over a number of queries. In operational settings, even single negative outliers can impact the user's perception of a system substantially. We have observed that combination approaches very effectively address the desire to avoid negative outliers. This is especially true for the combination of query translation (QT) and document translation (DT), which helps to boost queries that perform poorly when only using one method alone. Tables 3 and 4 show an analysis of the benefits gained by combining DT + QT as opposed to using either method alone. In all three campaigns, the combination experiment outperforms both DT and QT experiments. In CLEF 2002, however, we have observed that the run using only DT via MT performed practically identically when considering average precision to the more complicated combination run. However, looking at the recall, the combined run retrieves

Table 3. Average precision results for document translation (DT), query translation (QT) and a combination of both. Combination outperforms the simpler methods in all cases.

	CLEF 2000	CLEF 2001	CLEF 2002
DT	0.2816	0.3099	0.3539
QT	0.2500	0.2773	0.2876
DT + QT	0.3107	0.3416	0.3554

Table 4. Direct per-query comparison of document translation (DT), query translation (QT) and a combination of both. Shown is the number of queries performing better in terms of average precision for the given alternatives. For example, the combined run for the 2002 campaign outperforms the respective document translation run for 28 out of 50 queries.

	CLEF 2000	CLEF 2001	CLEF 2002
DT + QT vs. DT	32:8	41:9	28:22
DT + QT vs. QT	31:9	36:14	41:9
DT vs. QT	24:16	25:25	30:20

nearly 10% more relevant items. While not superior in terms of the popular average precision measure, the combined DT + QT run seems therefore preferable in operational situations.

In both CLEF 2000 and CLEF 2001, the picture is even more favorable for the combined run, which clearly outperforms both DT and QT alone. When looking at individual queries, one can see that combination is almost always beneficial, with the vast majority of queries showing improvement. Even more important, practically all those queries that show decreased performance suffer only slightly.

As a side note, the document translation runs consistently outperformed query translation runs for all three campaigns. However, when the merging process used for fusing the outputs from the different language pairs in query translation is investigated, it becomes apparent that much of this difference stems from merging. We will return to this aspect at the end of this section.

One of the measures published by CLEF is the performance compared to a theoretical median of all participants for every individual query. Our goal is therefore to outperform at least this median for as many cases as possible. This again allows us to conclude that we avoid producing negative outliers. In CLEF 2001, we observed that the number of queries with performance below the median decreased by as much as half when comparing a system combining both QT and DT with a system employing either of the two translation options alone (see Table 5).

5.3. Broad applicability of the system to a range of operational settings

We try to ensure the applicability of the system to different settings by avoiding any undue "overtuning" to the collection. The choice of a symmetrical handling of all language pairs

Table 5. Comparison of the number of queries above and below the median performance of all CLEF participants in terms of average precision, shown for systems using document translation (DT), query translation (QT) and a combination of both. For example, our document translation run for the 2000 campaign outperformed the median performance for 30 out of 40 queries.

	CLEF 2000	CLEF 2001	CLEF 2002
DT vs. median	30:10	24:26	38:10
QT vs. median	23:15	22:22	39:11
DT + QT vs. median	30:9	37:12	39:8

stems from this desire. We have observed in several cases that it would have been beneficial to chose a different approach for individual language pairs when focusing strictly on performance within the CLEF campaign. In CLEF 2000, for example, we used a dictionary-type word list to combine with machine translation for the German/English language pair. This additional resource hurt the overall result by 25% when compared to machine translation alone. Other examples include the English similarity thesaurus we used in both CLEF 2001 and CLEF 2002. This component hurt overall performance due to noisy training data, but still provided valuable improvements for some individual queries.

While we were therefore "ill-served" by the choice of including some of these resources for within-CLEF evaluation, we feel that we can conclude that the combination system was robust to such design. We experienced only slight performance degradation while still benefiting from the expanded lexical coverage that additional resources provide.

5.4. Evaluation of merging

The process of merging intermediate results is central to our combination approach, and merits evaluation outside the scope of the criteria outlined above. As observed in our discussion of negative outliers, document translation consistently outperforms query translation in all three campaigns (Table 3). However, when the performance of the query translation run is subdivided into individual retrieval performance and subsequent merging, we can in retrospect investigate through the use of the relevance assessments provided by CLEF the performance loss incurred by imperfect merging. The knowledge of which items are relevant to a query allows us to produce a "perfect" merging result by taking the optimal ratio of items from all intermediate results. By doing this, we can see that there remains a lot of potential in developing better merging strategies.

An optimally merged query translation run (average precision of 0.4876, see Table 6) clearly outperforms document translation (0.3539, see Table 3). However, all our merging strategies miss the optimal performance by a wide margin. Thus, the avoidance of the merging problem by the document translation run turns out to be a real advantage.

When looking at the individual methods, we see that taking collection sizes into account when interleaving gives a small benefit compared to straight-forward interleaving. This result is consistent for the various query lengths we have tried.

The merging strategy based on feedback shows little improvement compared to the simpler interleaving methods, even though it allows different merging ratios for different

Table 6. Performance of actual and theoretical merging strategies.

Merging strategy	CLEF 2002 MT query translation
Interleaving	0.3249
Collection size-based interleaving	0.3369
Feedback merging	0.3220
Relevance ratio	0.4067
Optimal	0.4876

queries. This is disappointing, since the use of variable merging ratios makes it possible to lift one of the main restrictions of the simpler merging methods. When looking at the ratios which were actually used for merging, we see that this method does indeed chose fairly different ratios for individual queries. However, we believe that these differences were obscured by the relatively large number of relevant items in the CLEF collections, and by problems with the estimation of the ratios from the feedback terms. While we saw encouraging improvements for some queries, feedback merging therefore did not bring any sustained improvement in retrieval effectiveness, even being slightly disadvantageous in several cases. We are working on improvements to this method.

The last two methods listed in Table 6 give theoretical upper bounds for merging performance. The "relevance ratio" experiment was obtained by determining the correct merging ratio per query based on the relevance assessments. This run shows the theoretical optimum which could be obtained by a method such as the feedback merging strategy. An even better performance is obtainable if different merging ratios are adopted not only per query, but also for different levels of recall within a query (the necessary merging strategy was demonstrated by Chen (2002b) during a presentation at the CLEF 2002 workshop). This upper bound is listed under the label "optimal" in Table 6. The huge performance difference between the actual merging strategies available and this optimum validates our belief that merging remains one of the big challenges in the design of better combination systems.

6. Conclusions

Eurospider has participated in all three CLEF evaluation campaigns to date. Our main objective was to test our CLIR component developed on the basis of a combination approach to CLIR. Our focus was on keeping the experiments applicable to operational settings. The goals of the combination approach in this respect were twofold: improved retrieval effectiveness, and improved system robustness with regard to negative outliers, lexical coverage and applicability to a broad range of domains. However, the second goal has not been explored extensively inside the CLEF campaign.

We have demonstrated how our component can combine a wide range of different aspects: different language pairs, different types of translation resources, and different levels of translation (document translation vs. query translation). While problems remain with merging, we have found definite benefits from a combination of multiple simpler approaches.

The combination of document translation and query translation improves performance for all three campaigns, in two cases substantially. The combined runs also perform better when compared query-by-query to the individual runs or the median CLEF performance.

Combination of multiple translation resources improves lexical coverage, and in some cases also retrieval performance. It also helps to avoid negative outliers, which is very important with respect to operational settings.

We have observed in all three campaigns that our experiments using only document translation outperform those using query translation alone. However, when investigating more closely, we find that this behavior is mainly due to a big drop in performance attributable to imperfect merging. Since merging is a central aspect of a combination system such as ours, with intermediate results merged across languages, translation resources and translation methods, this indicates that more work is needed on the merging problem, and that the combination approach has a lot of potential remaining for future improvements.

Acknowledgments

Peter Schäuble, Bärbel Ripplinger and Anne Göhring were involved in planning and conducting the Eurospider participation in the three CLEF campaigns. The paper has benefited from the work of three anonymous reviewers, who provided detailed and helpful comments. Their input is greatly appreciated.

Note

1. Incidentially, a small subfield of CLIR has since sprung up that tries to build systems that use no translation at all, adding a fourth dimension to this kind of classification, see e.g. McNamee and Mayfield (2002a).

References

Ballesteros L and Croft B (1996) Dictionary methods for cross-lingual information retrieval. In: Proceedings of the 7th International DEXA Conference on Database and Expert Systems, Sept. 9–13, Zurich, Switzerland. Springer, pp. 791–801.

Bartell BT, Cottrell GW and Belew RK (1994) Automatic combination of multiple ranked retrieval systems. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 173–181.

Belkin NJ, Cool C, Croft WB and Callan JP (1993a) The effect of multiple query representations on information retrieval system performance. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 339–346.

Belkin NJ, Kantor P, Cool C and Quatrain R (1993b) Combining evidence for information retrieval. In: The Second Text REtrieval Conference (TREC-2), NIST Special Publication 500-215, pp. 35–43.

Brand R and Brünner M (2002) Océ at CLEF 2002. In: Working Notes for the CLEF 2002 Workshop, pp. 21–30.

Braschler M, Krause J, Peters C and Schäuble P (1998a) Cross-language information retrieval (CLIR) track overview. In: The Seventh Text REtrieval Conference (TREC-7), NIST Special Publication 500-242, pp. 1–8.

Braschler M, Mateev B, Mittendorf E, Schäuble P and Wechsler M (1998b) SPIDER retrieval system at TREC7. In: The Seventh Text REtrieval Conference (TREC-7), NIST Special Publication 500-242, pp. 446–454.

Braschler M and Schäuble P (1998) Multilingual information retrieval based on document alignment techniques. In: Research and Advanced Technology for Digital Libraries, Second European Conference, ECDL '98, Lecture Notes in Computer Science, Vol. 1513, Springer, pp. 183–197.

Braschler M and Schäuble P (2001) Experiments with the eurospider retrieval system for CLEF 2000, CLEF 2000. In: Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lecture Notes in Computer Science, Vol. 2069, Springer, pp. 140–148.

Braschler M, Ripplinger B and Schäuble P (2002a) Experiments with the eurospider retrieval system for CLEF 2001. In: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406, Springer, 2002, pp. 102–110.

Braschler M, Göhring A and Schäuble P (2002b) Eurospider at CLEF 2002. In: Working Notes for the CLEF 2002 Workshop, pp. 127–132.

Callan JP, Lu Z and Croft WB (1995) Searching distributed collections with inference networks. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 21–28.

Chen A (2002a) Multilingual information retrieval using English and Chinese queries. In: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406, Springer, pp. 44–58.

Chen A (2002b) Cross-language retrieval experiments at CLEF 2002. In: Working Notes for the CLEF 2002 Workshop, pp. 5–20.

Du A and Callan J (1998) Probing a collection to discover its language model. Technical Report 98-29, Department of Computer Science, University of Massachusetts.

Figuerola CG, Berrocal JLA, Zazo AF and Díaz RG (2001) A simple approach to the Spanish-English bilingual retrieval task. In: Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lecture Notes in Computer Science, Vol. 2069, Springer, pp. 224–229.

Fox EA, Koushik MP, Shaw J, Modlin R and Rao D (1992) Combining evidence from multiple searches. In: The First Text REtrieval Conference (TREC-1), NIST Special Publication 500-207, pp. 319–328.

Fox EA and Shaw JA (1993) Combination of multiple searches. In: The Second Text REtrieval Conference (TREC-2), NIST Special Publication 500-215, pp. 243–252.

Gey F, Jiang H, Petras V and Chen A (2001) Cross-language retrieval for the CLEF collections—comparing multiple methods of retrieval. In: Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lecture Notes in Computer Science, Vol. 2069, Springer, pp. 116–128.

Gey FC, Jiang H and Perelman N (2002) Working with Russian queries for the GIRT, bilingual and multilingual CLEF tasks. In: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406, Springer, pp. 235–243.

Grefenstette G (1998) Cross-Language Information Retrieval. Kluwer Academic Publishers.

Hedlund T, Keskustalo H, Pirkola A, Sepponen M and Järvelin K (2001) Bilingual tests with Swedish, Finnish, and German queries: Dealing with morphology, compound words, and query structure. In: Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lecture Notes in Computer Science, Vol. 2069, Springer, pp. 210–223.

Hiemstra D, Kraaij W and Pohlmann R (2001) Translation resources, merging strategies, and relevance feedback for cross-language information retrieval. In: Cross-Language Information Retrieval and Evaluation, Workshop of the Cross-Language Evaluation Forum, CLEF 2000, Lecture Notes in Computer Science, Vol. 2069, Springer, pp. 102–115.

Jones GJF and Lam-Adesina AM (2002) Exeter at CLEF 2001: Experiments with machine translation for bilingual retrieval. In: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406, Springer, pp. 59–77.

Lee JH (1995) Combining multiple evidence from different properties of weighting schemes. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 180–188.

McNamee P and Mayfield J (2002a) JHU/APL experiments at CLEF: Translation resources and score normalization. In: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406, Springer, pp. 193–208.

McNamee P and Mayfield J (2002b) Scalable multilingual information access. In: Working Notes for the CLEF 2002 Workshop, pp. 133–140.

Moulinier I and Molina-Salgado H (2002) Thomson legal and regulatory experiments for CLEF 2002. In: Working Notes for the CLEF 2002 Workshop, pp. 91–96.

Nie JY and Simard M (2002) Using statistical translation models for bilingual IR. In: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406, Springer, pp. 137–150.

Oard DW (1997) Alternative approaches for crosslanguage text retrieval. In: AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for Artificial Intelligence.

Qiu Y and Frei HP (1993) Concept based query expansion. In: Proceedings of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 160–169.

Rogati M and Yang Y (2002) Cross-lingual pseudo-relevance feedback using a comparable corpus. In: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406, Springer, pp. 151–157.

Salton G (1970) Automatic processing of foreign language documents. Journal of the American Society for Information Science, 21:187–194.

Saracevic T and Kantor P (1988) A study of information seeking and retrieving. III. Searchers, searches and overlap. Journal of the American Society for Information Science, 39:197–216.

Savoy J (2002a) Report on CLEF-2001 experiments: Effective combined query-translation approach. In: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406, Springer, pp. 27–43.

Savoy J (2002b) Report on CLEF-2002 experiments: Combining multiple sources of evidence. In: Working Notes for the CLEF 2002 Workshop, pp. 31–46.

Sheridan P, Braschler M and Schäuble P (1997) Cross-language information retrieval in a multilingual legal domain. In: Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries, pp. 253–268.

Singhal A, Buckley C and Mitra M (1996) Pivoted document length normalization. In: Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 21–29.

Voorhees EM, Gupta NK and Johnson-Laird B (1995) Learning collection fusion strategies. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 172–179.