Published in Lecture Notes in Computer Science 3491, 27-37, 2005
which should be used for any reference to this work

Selection and Merging Strategies for Multilingual Information Retrieval

Jacques Savoy and Pierre-Yves Berger

Institut interfacultaire d'informatique, Université de Neuchâtel, Pierre-à-Mazel 7, 2001 Neuchâtel, Switzerland
{Jacques.Savoy, Pierre-Yves.Berger}@unine.ch

Abstract. In our fourth participation in the CLEF evaluation campaigns, our objective was to verify whether our combined query translation approach would work well with new requests and new languages (Russian and Portuguese in this case). As a second objective, we were to suggest a selection procedure able to extract a smaller number of documents from collections that seemed to contain no or only a few relevant items for the current request. We also applied different merging strategies in order to obtain more evidence about their respective relative merits.

1 Introduction

Based on our bilingual and multilingual experiments of the last years [1], [2], we conducted additional experiments involving various bilingual and multilingual test-collections. Based on a request written in English, we retrieved documents written in English, French, Finnish and Russian. As with previous experiments [2], we adopted a combined query translation strategy capable of submitting queries to documents written in various European languages, based on an original request written in English. Once the query translation phase was completed, we searched in the corresponding document collection using our retrieval scheme (bilingual). In Section 3, we carried out multilingual information retrieval, investigating various merging strategies based on the results obtained during our bilingual searches.

2 Bilingual Information Retrieval

In our experiments, we chose English as the language for submitting queries to be automatically translated into four different languages, using nine different machine translation (MT) systems and one bilingual dictionary ("Babylon"). The following freely available translation tools were used in our experiments:

1. SYSTRAN www.systranlinks.com
2. GOOGLE www.google.com/language_tools
3. FREETRANSLATION www.freetranslation.com
4. INTERTRAN intertran.tranexp.com/

5. Reverso Online www.reverso.fr/url_translation.asp
6. WorldLingo www.worldlingo.com/
7. BabelFish babelfish.altavista.com/
8. Prompt webtranslation.paralink.com/
9. Online www.online-translator.com/
10. Babylon www.babylon.com.

When using the Babylon bilingual dictionary to translate an English request word-by-word, usually more than one translation is provided, in an unspecified order. We decided to pick only the first translation available (labeled "Babylon 1"), the first two terms (labeled "Babylon 2") or the first three available translations (labeled "Babylon 3").

Table 1. Mean average precision of various single translation devices (TD queries, Okapi model)

	Mean average precision (% of monolingual search)			
TD queries Index	French word 49 queries	Finnish 4-gram 45 queries	Russian word 34 queries	Portuguese word 46 queries
Manual	0.4685	0.5385	0.3800	0.4835
Systran	0.3729 (79.6%)	N/A	0.2077 (54.7%)	0.3329 (68.9%)
Google	0.3680 (78.5%)	N/A	N/A	0.3375 (69.8%)
FreeTrans.	**0.3845** (82.1%)	N/A	**0.3067** (80.7%)	**0.4057** (83.9%)
InterTran	0.2664 (56.9%)	**0.2653** (49.3%)	0.1216 (32.0%)	0.3277 (67.8%)
Reverso	0.3830 (81.8%)	N/A	N/A	N/A
WorldLingo	0.3728 (79.6%)	N/A	0.2077 (54.7%)	0.3311 (68.5%)
BabelFish	0.3729 (79.6%)	N/A	0.2077 (54.7%)	0.3329 (68.9%)
Prompt	N/A	N/A	0.2960 (77.9%)	N/A
Online	N/A	N/A	0.2888 (76.0%)	0.3879 (80.2%)
Babylon 1	0.3706 (79.1%)	0.1965 (36.5%)	0.2209 (58.1%)	0.3071 (63.5%)
Babylon 2	0.3356 (71.6%)	N/A	0.2245 (59.1%)	0.2892 (59.8%)
Babylon 3	0.3378 (72.1%)	N/A	0.2243 (59.0%)	0.2858 (59.1%)

Table 1 shows the mean average precision obtained using the various translation tools and the Okapi probabilistic model (see [3] for implementation details). Of course, not all tools can be used for each language, and thus as shown in Table 1, various entries are missing (indicated with the label "N/A"). From this data, we can see that the results from the FreeTranslation MT system usually obtain satisfactory retrieval performances (around 82% of the mean average precision obtained by the corresponding monolingual search). As another good translation systems, we found that Reverso, BabelFish or WorldLingo worked well for French, Prompt for Russian or Online for both the Russian and Portuguese languages. For Finnish we found only two translation tools, but unfortunately their overall performance levels were not very good (similar to low

Table 2. Mean average precision of various combined translation devices (Okapi)

	Mean average precision			
TD queries Index Model	French word 49 queries	Finnish 4-gram 45 queries	Russian word 34 queries	Portuguese word 46 queries
Comb 1	Bab2+Free	Bab1+Inter	Bab1+Free	Free+Online
Comb 2	Bab2+Reverso		Free+Prompt	Bab1+Systran
Comb 3	Reverso+Systran		Prompt+Online	Bab1+Free+Onl
Comb 4	Free+Reverso		Free+Online	Bab1+Free+Sys
Comb 5	Bab2+Free+ Reverso		Bab1+Free+ Online	Bab1+Free+ Online+Systran
Best single	0.3845	0.2653	0.3067	0.4057
Comb 1	0.3784	**0.3042**	**0.3888**	0.4072
Comb 2	0.3857		0.3032	0.3713
Comb 3	0.3858		0.2964	**0.4204**
Comb 4	**0.4066**		0.3043	0.3996
Comb 5	0.3962		0.3324	0.4070

Table 3. Mean average precision of automatically translated queries (without automatic query expansion)

	Mean average precision			
TD queries Index Model	French word 49 queries Comb 4	Finnish 4-gram 45 queries Comb 1	Russian word 34 queries Comb 1	Portuguese word 46 queries Comb 3
Okapi	0.4066	**0.3042**	**0.3888**	**0.4204**
Prosit	0.4111	0.2853	0.3050	0.4085
Round-robin	**0.4129**	0.2969	0.3237	0.4129
Sum RSV	0.4111	0.2965	0.3707	0.4134
Norm Max	0.4096	0.2936	0.3610	0.4152
Norm RSV (Eq. 1)	0.4102	0.2937	0.3617	0.4152
Z-score (Eq. 3)	0.4098	0.2937	0.3618	0.4152
Z-scoreW (Eq. 3)	0.4100	0.2965	0.3645	0.4043

performance levels found when translating English topics into various Asian languages [4]). Not surprisingly we found that there were certain similarities and dissimilarities between the various translation tools. For example, the Systran, BabelFish, and WorldLingo MT systems appeared to be nearly identical MT systems.

To determine whether or not a given search strategy was better than another, we developed a decision rule. This was based on statistical validation using the bootstrap approach [5]. Thus, in the tables presented in this paper we underlined statistically significant differences based on a two-sided non-parametric boot-

Table 4. Mean average precision of automatically translated queries (after blind query expansion)

	Mean average precision			
TD queries	French	Finnish	Russian	Portuguese
Index	word	4-gram	word	word
	49 queries	45 queries	34 queries	46 queries
Model	Comb 4	Comb 1	Comb 1	Comb 3
Okapi (#d/#t)	0.4197 (5/15)	0.3225 (5/150)	0.3888 (0/0)	0.4373 (10/75)
Prosit (#d/#t)	0.4251 (10/15)	0.2960 (5/40)	**0.3945 (5/20)**	0.4805 (10/30)
Round-robin	0.4275	**0.3308**	0.3152	0.4767
Sum RSV	0.4307	0.2970	0.3713	**0.4854**
Norm Max	0.4320	0.3035	0.3174	0.4815
Norm RSV (Eq. 1)	0.4325	0.3041	0.3139	0.4788
Z-score (Eq. 3)	0.4323	0.3001	0.3068	0.4840
Z-scoreW (Eq. 3)	**0.4330**	0.3007	0.3088	0.4851

Table 5. Description and mean average precision of our official bilingual runs

	Russian	Russian	Portuguese	Portuguese
	34 queries	34 queries	46 queries	46 queries
IR 1 (#d/#t)	Prosit (3/15)	Prosit (3/15)	Prosit (10/20)	Okapi (0/0)
IR 2 (#d/#t)	Okapi (3/15)	Okapi (3/10)	Okapi (5/15)	Prosit (0/0)
Data fusion	Round-robin	Round-robin	Norm RSV	Norm RSV
Translation	Free-Reverso	Pro-Free-Rever	Onl-Free-Bab1	Onl-Free-Sys-Bab1
MAP	**0.3007**	0.2962	**0.4704**	0.4491
Run name	UniNEBru1	UniNEBru2	UniNEBpt1	UniNEBpt2

strap test, for any means that had a significance level fixed at 5%. As shown in Table 1, we used the best translation system (depicted in bold) as the baseline. As depicted, differences in mean average precision between the manually translated queries and the best automatic translation tools are always statistically significant, except for the Russian collection. On the other hand, differences between the various translation tools are usually not statistically significant, except for a few such as "Babylon 2" and "Babylon 3" for both French and Portuguese, or "InterTran" for French and Russian.

It is known that although a given translation tool may produce acceptable translations for a given set of requests, it may perform poorly for other queries [1], [2]. To date we have not been able to detect with much precision when a given translation will produce satisfactory retrieval performance and when it will fail. In this vein, Kishida *et al.* [6] suggest using a linear regression model to predict the average precision of the current query, based on both manual evaluations of translation quality and the underlying topic difficulty.

In order to hopefully improve retrieval performance, in this study we chose to concatenate two or more translations before submitting a query for translation.

Table 2 shows the retrieval effectiveness for such combinations, using the Okapi probabilistic model. The top part of the table indicates the exact query translation combination used while the bottom part shows the mean average precision achieved by our combined query translation approach. When selecting the query translations to be combined, a priori we considered the best translation tools.

The resulting retrieval performances shown in Table 2 are sometimes better than the best single translation scheme, as indicated in the row labeled "Best single" (e.g., the strategies "Comb 4" or "Comb 5" for French, or "Comb 1" for Russian, and "Comb 3" for Portuguese). Statistically however none of these combined query translation approaches performs better than the best single translation tool.

Of course, the main difficulty in this bilingual search was the translation of English topics into Finnish, due to the limited number of free translation tools available. When handling any languages from around the world that are less frequently used, it seems it would be worthwhile considering other translation alternatives, such as probabilistic translation based on parallel corpora [7], [8].

As described in [3], for monolingual searches we used a data fusion search strategy that combined the Okapi and Prosit probabilistic models. As shown in Table 3, in the current context our data fusion approaches do not improve retrieval effectiveness. However, differences in mean average precision are usually not statistically significant, except for the Finnish corpus where all data fusion approaches used significantly decrease retrieval performance.

Of course before combining the result lists we could also automatically expand the translated queries, using a pseudo-relevance feedback method (Rocchio's approach in the present case). As shown in Table 4, the resulting mean average precision after combining the two IR models (after pseudo-relevance feedback) did not always improve retrieval effectiveness, when compared to the best single approach. Moreover, the statistical tests did not reveal any significant differences. In Tables 3 and 4, under the heading "Z-scoreW", we attached a weight of 1.5 to the best single IR model, and 1 to the other.

Finally, Table 5 lists the parameter settings used for our official runs in the bilingual task. Each experiment uses queries written in English to retrieve documents written either in Russian or in Portuguese.

3 Multilingual Information Retrieval

Our multilingual information retrieval system is based on the use of a query translation strategy instead of either translating all documents into a common language (e.g., English), combining both query and document translations [9] or ignoring the translation phase [10], [8]. For a general overview of these issues, see [11]). In our approach, when a request was received (in English in this study), we automatically translated it into the desired target languages and then searched for pertinent items within each of the four corpora (English, French, Finnish and Russian). We then applied a merging procedure to take each result list received from the search engines, thus providing a single ranked result. As a

first solution to this procedure, we considered the round-robin approach whereby we took one document in turn from each individual list [12].

To account for the document score computed for each retrieved item (denoted RSV_k for document D_k), 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 [13]. Such a strategy is called raw-score merging and produces a final list sorted by the document score computed by each collection. When using the same IR model (with the same or very similar parameter settings) to search all collections, such a merging strategy should result in good retrieval performance (e.g., with a logistic regression IR model in [14]).

Unfortunately, the document scores cannot always be directly compared and thus we introduced a third merging strategy by normalizing the document scores within each collection. This was done by dividing the scores by the maximum score (i.e. the document score of the retrieved record in the first position) and denoted them "Norm Max". As a variant of this normalized score merging scheme (denoted "Norm RSV"), we could normalize the document RSV_k scores within the ith result list, according to the following formula:

$$Norm\ RSV_k\ =\ \frac{RSV_k\ -\ MinRSV^i}{MaxRSV^i\ -\ MinRSV^i} \tag{1}$$

As a fifth merging strategy, we might use logistic regression to predict the probability of a binary outcome variable, according to a set of explanatory variables [15]. In our current case, we predicted the probability of relevance for document D_k, given both the logarithm of its rank (indicated by $ln(Rank_k)$) and the original document score RSV_k as indicated in Equation 2. Based on these estimated relevance probabilities (computed independently for each language using S+ software), we sorted the records retrieved from separate collections in order to obtain a single ranked list. This approach requires that a training set is available, in order to estimate the underlying parameters. To achieve this, we used the CLEF-2003 topics and their relevance assessments in our evaluations.

$$Prob\,[D_k\ is\ rel\ |\ Rank_k,\ RSV_k]\ =\ \frac{e^{\alpha+\beta_1\cdot ln(Rank_k)+\beta_2\cdot RSV_k}}{1+e^{\alpha+\beta_1\cdot ln(Rank_k)+\beta_2\cdot RSV_k}} \tag{2}$$

As a final strategy we suggest merging the retrieved documents according to the Z-score, calculated on the basis of their document scores [2]. Within this scheme, for the ith result list, we needed to compute average for the RSV_k (denoted μRSV^i) and the standard deviation (denoted σRSV^i). Based on these values, we can normalize the retrieval status value of each document D_k provided by the ith result list, by applying the following formula:

$$Zscore\ RSV_k\ =\ \alpha_i\cdot\left[\frac{RSV_k-\mu RSV^i}{\sigma RSV^i}+\delta_i\right]\ \ \delta_i=\frac{\mu RSV_k-MinRSV^i}{\sigma RSV^i} \tag{3}$$

where the value of δ^i is used to generate only positive values, and α_i (usually fixed at 1) is used to reflect the retrieval performance of the underlying retrieval model

and to account for the fact that pertinent items are not uniformly distributed across all collections.

Table 6 lists the exact parameters used to query the four different collections. For the Russian collection, we only considered the word-based indexing strategy while for the Finnish language we only used the 4-gram indexing scheme. The top part of Table 6 shows how we used a combined query translation strategy for French, Finnish and Russian languages (Condition A). As described in our monolingual experiments [3], we might also apply a data fusion phase before merging the result lists. Thus, when searching the English or French corpus, we combined the Okapi and Prosit result lists (both with blind query expansion). In a second multilingual experiment (denoted Condition B), we applied a data fusion approach for all bilingual searches (descriptions given in the middle part of Table 6). Finally, we decided to search through all corpora using the same retrieval model, Prosit in this case, as shown in the bottom part of Table 6 (and corresponding to Condition C).

Table 7 lists the retrieval effectiveness of various merging strategies using three different bilingual search parameter settings. In this table, the round-robin scheme was used as a baseline. On the one hand, when different search engines were merged (Condition A and Condition B), the raw-score merging strategy resulted in very poor mean average precision and differences with the round-robin approach are statistically significant. On the other hand, when the same search engine is used (Condition C), the resulting performance of the raw score merg-

Table 6. Description of various runs done separately on each corpus (descriptions listed at top form Condition A, the middle Condition B, and bottom Condition C)

	Parameters of each single run according to each language			
TD queries	English 42 queries	French 49 queries	Finnish (4-gram) 45 queries	Russian (word) 34 queries
Condition A				
IR 1 (#d/#t)	Okapi (3/15)	Prosit (5/15)	Okapi (5/30)	Prosit (3/15)
IR 2 (#d/#t)	Prosit (3/10)	Okapi (5/10)		
Data fusion	Z-score	Z-scoreW		
Translation		Bab2-Free-Rev	Bab1-Inter	Rev-Free
MAP	0.5580	0.4098	0.2956	0.2914
Condition B				
IR 1 (#d/#t)	Okapi (3/15)	Prosit (5/15)	Okapi (5/30)	Prosit (3/15)
IR 2 (#d/#t)	Prosit (3/10)	Okapi (5/10)	Lnu-ltc (3/40)	Okapi (3/15)
Data fusion	Z-score	Z-scoreW	Round-robin	Round-robin
Translation		Bab2-Free-Rev	Bab1-Inter	Rev-Free
MAP	0.5580	0.4098	0.3080	0.3007
Condition C				
IR (#d/#t)	Prosit (3/10)	Prosit (5/15)	Prosit (10/30)	Prosit (3/15)
Translation		Bab2-Fre-Rev	Bab1-Inter	Rev-Free
MAP	0.5633	0.4055	0.2909	0.2914

Table 7. Mean average precision of various merging strategies (TD queries)

	Mean average precision (% change)		
Parameter setting Merging Strategy	Condition A 50 queries	Condition B 50 queries	Condition C 50 queries
Round-robin (baseline)	0.2386	0.2430	0.2358
Raw-score	0.0642 (-73.1%)	0.0650 (-73.2%)	0.3067 (+30.1%)
Norm Max	0.2552 (+7.0%)	0.1044 (-57.0%)	0.2484 (+5.3%)
Norm RSV (Eq. 1)	0.2899 (+21.5%)	0.1042 (-57.1%)	0.2646 (+12.2%)
Log. reg. (ln(rank),RSV)	**0.3090** (+29.5%)	**0.3111** (+28.0%)	**0.3393** (+43.9%)
Biased round-robin	0.2639 (+10.6%)	0.2683 (+10.4%)	0.2613 (+10.8%)
Z-score (Eq. 3)	0.2677 (+12.2%)	0.2903 (+19.5%)	0.2555 (+8.4%)
Z-score (Eq. 3) α_i=1.5	0.2669 (+11.9%)	0.3019 (+24.2%)	0.2867 (+21.6%)
Log. reg. & Select. (0)	0.2957 (+23.9%)	0.2959 (+21.8%)	0.3405 (+44.4%)
Log. reg. & Select. (3)	0.2953 (+23.8%)	0.2982 (+22.7%)	0.3378 (+43.3%)
Log. reg. & Select. (10)	0.2990 (+25.3%)	0.3008 (+23.8%)	0.3381 (+43.4%)
Log. reg. & Select. (20)	0.3010 (+26.1%)	0.3029 (+24.7%)	0.3384 (+43.5%)
Log. reg. & Select. (50)	0.3044 (+27.6%)	0.3064 (+26.1%)	0.3388 (+43.7%)
Log. reg. & OptSelect.	**0.3234** (+35.5%)	**0.3261** (+34.2%)	**0.3558** (+50.9%)

ing is statistically better than the baseline. Normalized score merging based on Equation 1 results in statistically significant degradation compared to the simple round-robin approach when using the parameter settings of Condition B (0.1042 vs. 0.2430, or -57.2% in relative performance). By applying our logistic model using both the rank and the document score as explanatory variables, the resulting mean average precision is statistically better than the round-robin merging strategy, and better than the other merging approaches. Under Condition B however, the difference between our logistic model and the Z-score merging strategy is rather small (0.3111 vs. 0.3019, or 3% in relative performance).

As a simple alternative, we could also suggest a biased round-robin approach which extracts not one document per collection per round, but one document for the Russian corpus and two from the English, French and Finnish collections (because the last three represent larger corpora). This merging strategy provides good retrieval performance, better that of the simple round-robin approach. Finally, the Z-score merging approach seems result in generally satisfactory performance. Moreover, we may multiply the Z-score by an α value (performance under the label "$\alpha_i = 1.5$" where the α_i values set as follows: EN: 1.5, FR: 1.5, FI: 1.0, and RU: 1.0).

It cannot be expected however that each result list would always contain pertinent items, in response to a given request. In fact, a given corpus may contain no relevant information regarding the submitted request or the pertinent articles could not be found by the search engine. In the cross-lingual environment we discovered an additional problem: important facets of the original request were translated with inappropriate words or expressions. In all these cases, it is not useful to include items provided by such collections (or such search engines) in the final result list. In addition, the number of pertinent documents is usually

Table 8. Description and mean average precision of our official automatic multilingual runs

Run name	Query	Language	Merging	Parameters	MAP
UniNEmulti1	TD	English	Logistic	Condition A	**0.3090**
UniNEmulti2	TD	English	Z-scoreW	Cond. A, $\alpha_i = 1.5$	0.2969
UniNEmulti3	TD	English	raw-score	Condition C	0.3067
UniNEmulti4	TD	English	Log. & select	Cond. A, $m = 20$	0.3010
UniNEmulti5	TD	English	Z-scoreW	Condition B	0.3019

not uniformly distributed across all four collections. For a given request (e.g., related to a regional or a national event), only one or two collections may contain relevant documents describing this particular event.

To account for these phenomena, we designed a selection procedure that works as follows. First, for each result list we normalize the document score according to our logistic regression method (given in Equation 2). After this step, each document score represents the probability that the underlying article is relevant (with respect to the query submitted and the collection). It is also interesting to note that these probabilities are obtained after a blind query expansion and therefore the number of search terms are more or less the same across queries.

In the second step, for each result list (or language), we sum the document scores of the top 15 ranked documents. If this sum exceeded a given threshold (depending on the collection or search engine), we could thus assume that the corresponding collection contained many pertinent documents. Otherwise, we might only include the m best ranking retrieved items from the corpus (with a relatively small m value). This allows us to limit the number of items extracted from a given corpus while also taking account of the fact that each collection usually contains few pertinent items. Table 7 lists the mean average precision achieved using this selection strategy under the label "Log. reg. & Select. (m)," where the value m indicates that we always include the m best retrieved items from each corpus in our final result list. Of course, when we set $m = 0$, the system will not extract any documents from a collection having a poor overall score. Finally under the label "Log. reg. & OptSelect.", we computed the mean average precision that could be achieved when selection occurs without any errors (with $m = 0$). When using such an ideal selection system, the mean average precision is clearly better than all other merging strategies (e.g, under Condition C, the mean average precision is 0.3558 vs. 0.3393 with the logistic regression without selection).

Table 8 contains the descriptions of our official runs for the multilingual tracks. In the row entitled "UniNEmulti3", all searches were based on the Prosit retrieval model in order to obtain more comparable document score across the various collections. In this context, the raw-score merging strategy provides good overall performance levels.

4 Conclusion

In this fifth CLEF evaluation campaign, we assessed various query translation tools (see Table 1) used together with a combined translation strategy (see Table 2), that usually resulted in better levels of retrieval performance. However, the differences between the best single query translation tool and the various combinations of query translation strategies were not statistically significant. On the other hand, while a bilingual search can be viewed as easier for some language pairs (e.g., English query of a French document collection, or English of a Portuguese), this task is clearly more complex for other language pairs (e.g., English to Finnish). From combining various result lists (see Table 3 or 4), we cannot always obtain better retrieval effectiveness, where compared to isolated runs and the differences with the best single IR model are usually not statistically significant.

In multilingual tasks, searching documents written in different languages represents a real challenge. In this case we proposed a new simple selecting strategy which would avoid extracting a relatively large number of documents from collections containing many documents seeming to have no or little interest with respect to the current query (see Table 7). In this multilingual task, it was also interesting to mention that combining the result lists provided by the same search engine (Condition C in Table 7) may sometimes produce good retrieval effectiveness, as compared to combining different search models (Condition A in Table 7). If in our implementation combining different IR models did not present a statistically significant difference (see Table 4 and evaluations under Condition B in Table 7), the best multilingual system [16] of this evaluation campaign would be based on this combining approach.

Acknowledgments. The author would like to also thank the CLEF-2004 task organizers for their efforts in developing various European language test-collections. The author would also like to thank C. Buckley from SabIR for giving us the opportunity to use the SMART system, together with Samir Abdou for his help in translating the English topics. This research was supported by the Swiss National Science Foundation under Grant #21-66 742.01.

References

1. Savoy, J.: Combining Multiple Strategies for Effective Monolingual and Cross-Lingual Retrieval. IR Journal, **7** (2004) 121–148
2. Savoy, J.: Report on CLEF-2003 Multilingual Tracks. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (Eds.): Comparative Evaluation of Multilingual Information Access Systems. Lecture Notes in Computer Science 3237. Springer, Heidelberg (2004)
3. Savoy, J.: Data Fusion for Effective European Monolingual Information Retrieval. In this volume.
4. Savoy, J.: Report on CLIR task for the NTCIR-4 Evaluation Campaign. In Proceedings NTCIR-4. Tokyo (2004) 178–185

5. Savoy, J.: Statistical Inference in Retrieval Effectiveness Evaluation. Information Processing & Management, **33** (1997) 495–512
6. Kishida, K., Kuriyama, K., Kando, N., Eguchi, K.: Prediction of Performance on Cross-Lingual Information Retrieval by Regression Models. In Proceedings NTCIR-4. Tokyo (2004) 219–224
7. Nie, J.Y., Simard, M., Isabelle, P., Durand, R.: Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. In Proceedings of the ACM-SIGIR'99. The ACM Press, New York (1993) 74–81
8. MacNamee, P., Mayfield, J.: JHU/APL Experiments in Tokenization and Non-Word Translation. In: Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (Eds.): Comparative Evaluation of Multilingual Information Access Systems. Lecture Notes in Computer Science 3237. Springer, Heidelberg (2004)
9. Chen, A., Gey, F.: Combining Query Translation and Document Translation in Cross-Language Retrieval. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (Eds.): Advances in Cross-Language Information Retrieval. Lecture Notes in Computer Science. Springer, Heidelberg (2004), to appear
10. Buckley, C., Mitra, M., Waltz, J., Cardie, C.: Using Clustering and Superconcepts within SMART. In Proceedings TREC-6. NIST Publication #500-240, Gaithersburg (1998) 107–124
11. Braschler, M., Peters, C.: Cross-Language Evaluation Forum: Objectives, Results and Achievements. IR Journal, **7** (2004) 7–31
12. Voorhees, E.M., Gupta, N.K., Johnson-Laird, B.: The Collection Fusion Problem. In Proceedings TREC-3. NIST Publication #500-225, Gaithersburg (1995) 95–104
13. Kwok, K.L., Grunfeld, L., Lewis, D.D.: TREC-3 Ad-hoc, Routing Retrieval and Thresholding Experiments using PIRCS. In Proceedings TREC-3. NIST Publication #500-225, Gaithersburg (1995) 247–255
14. Chen, A.: Cross-language Retrieval Experiments at CLEF 2002. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (Eds.): Advances in Cross-Language Information Retrieval. Lecture Notes in Computer Science: Vol. 2785. Springer, Heidelberg (2003), 28–48
15. Le Calvé, A., Savoy, J.: Database Merging strategy based on Logistic Regression. Information Processing & Management, **36** (2000) 341–359
16. Adafre, S.F., van Hage, W.R., Kamps, J., de Melo, G.L., de Rijke, M.: The University of Amsterdam at at CLEF 2004. In this volume.