Selection and Merging Strategies for Multilingual Information Retrieval

Jacques Savoy, Pierre-Yves Berger

Institut interfacultaire d'informatique
Université de Neuchâtel, Switzerland
{Jacques.Savoy, Pierre-Yves.Berger}@unine.ch Web site: www.unine.ch/info/clef/

Abstract. For 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 suggested a selecting procedure that could extract a smaller number of documents from collections that for the current request seem to contain no or only few relevant items. We also applied different merging strategies in order to obtain more evidence on the respective relative merits.

Introduction

Based on our bilingual and multilingual experiments of last years (Savoy 2003; 2004a), we conducted different experiments involving various bilingual and multilingual test-collections. In the latter case, we retrieved documents written in the English, French, Finnish and Russian languages, based on a request written in English. As of last years, we adopted a combined query translation strategy that is able to produce queries in different European languages based on an original request written in English. Once the translation phase was completed, we searched in the corresponding document collection using our retrieval scheme (bilingual retrieval). In Section 2, we carried out multilingual information retrieval, investigating various merging strategies based on the results obtained during our bilingual searches.

1. Bilingual Information Retrieval

In our experiments, we chose English as the language to be used when submitting requests for automatic translation 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:

SYSTRAN	www.systranlinks.com/
GOOGLE	www.google.com/language_tools
FREETRANSLATION	www.freetranslation.com/web.htm
INTERTRAN	intertran.tranexp.com/
REVERSO	www.reverso.fr/url_translation.asp
WORLDLINGO	www.worldlingo.com/
BABELFISH	babelfish.altavista.com/
PROMPT	webtranslation.paralink.com/
ONLINE	www.online-translator.com/srvurl.asp?lang=en
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 shows the resulting mean average precision using the various translation tools and the Okapi probabilistic model (see Savoy (2004c) 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 MAP of the corresponding monolingual search). As another good translation system, we may mention Reverso or BabelFish for the French, Prompt for the Russian or Online for both the Russian and Portuguese languages. For the Finnish language we found only two translation tools, but unfortunately their overall performance levels were not very good (a similar low level performance was also found when translating English topics into various Asian languages (Savoy 2004b)). Not surprisingly, we found there was a relationship between the various translation tools. For example, the Systran, BabelFish, and WorldLingo MT systems appeared to be nearly identical MT systems.

Language	Mean average precision (% of monolingual search)				
	French Okapi 49 queries	Finnish Okapi (word) 45 queries	Finnish Okapi (4-gram) 45 queries	Russian Okapi (word) 34 queries	Portuguese Okapi 46 queries
Manual	0.4685	0.4773	0.5386	0.3800	0.4835
Systran	0.3729 (79.6%)	N/A	N/A	0.2077 (54.7%)	0.3329 (68.9%)
Google	0.3680 (78.5%)	N/A	N/A	N/A	0.3375 (69.8%)
FreeTrans	**0.3845** (82.1%)	N/A	N/A	**0.3067** (80.7%)	**0.4057** (83.9%)
InterTrans	0.2664 (56.9%)	**0.2290** (48.0%)	**0.2653** (49.3%)	0.1216 (32.0%)	0.3277 (67.8%)
Reverso	0.3830 (81.8%)	N/A	N/A	N/A	N/A
WorldLingo	0.3728 (79.6%)	N/A	N/A	0.2077 (54.7%)	0.3311 (68.5%)
BabelFish	0.3729 (79.6%)	N/A	N/A	0.2077 (54.7%)	0.3329 (68.9%)
Prompt	N/A	N/A	N/A	0.2960 (77.9%)	N/A
Online	N/A	N/A	N/A	0.2888 (76.0%)	0.3879 (80.2%)
Babylon 1	0.3706 (79.1%)	0.1771 (37.1%)	0.1965 (36.5%)	0.2209 (58.1%)	0.3071 (63.5%)
Babylon 2	0.3356 (71.6%)	N/A	N/A	0.2245 (59.1%)	0.2892 (59.8%)
Babylon 3	0.3378 (72.1%)	N/A	N/A	0.2243 (59.0%)	0.2858 (59.1%)

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

It is known that while a given translation tool may produce acceptable translations for a given set of requests, it may perform poorly for other queries (Savoy 2003; 2004a). To date we have not been able to detect very precisely when a given translation will produce satisfactory retrieval performance and when it will fail. In this vein, Kishida *et al*. (2004) suggest using a linear regression model to predict the average precision of the current query, based on both manual evaluations of translation quality for the current query and the underlying topic difficulty. In this study, before carrying out the retrievals, we chose to concatenate two or more translations before submitting a query for translation.

Language	Mean average precision				
	French Okapi 49 queries	Finnish Okapi (word) 45 queries	Finnish Okapi (4-gram) 45 queries	Russian Okapi (word) 34 queries	Portuguese Okapi 46 queries
Combination					
Comb 1	Bab2+Free	Bab1+Inter	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+Rev			Free+Online	Bab1+Free+Sys
Comb 5	Bab2+Free+ Reverso			Bab1+Free+ Online	Bab1+Free+ Online+Systran
Best single	0.3845	0.2290	0.2653	0.3067	0.4057
Comb 1	0.3784	**0.2529**	**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 2: Mean average precision of various combined translation devices (TD queries, Okapi model)

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 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 the Portuguese language). Of course, the main difficulty in this bilingual search was the translation of English topics into Finnish, due to limited number of free translation tools. When handling those languages less-often speaking around the world, it seems it would be worthwhile considering other translation alternatives, such as probabilistic translation based on parallel corpora (Nie *et al*. 1999), (MacNamee & Mayfield 2003).

For monolingual searches, as described in Savoy (2004c), we used a data fusion search strategy that combined the Okapi and Prosit probabilistic models (see details in Section 2). The data shown in Table 3 indicates that our data fusion approaches may result in better retrieval effectiveness (except for the Finnish 4-

gram indexing scheme or the Russian corpus). 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). The resulting mean average precision as shown in Table 4 did not improve the retrieval effectiveness when compared to the best single approach. In Tables 3 and 4, under the heading "Z-scoreW", we attached a weight of 1.5 to the Prosit model, and 1 to the Okapi model. Finally, Table 5 depicts the parameters used for our official bilingual runs.

	Mean average precision			
Language	French word 49 queries	Finnish 4-gram 45 queries	Russian word 34 queries	Portuguese word 46 queries
Combination \ Translation	Comb5	Comb1	Comb4	Comb3
Okapi	0.3962	**0.3042**	**0.3043**	**0.4204**
Prosit	0.3937	0.2853	0.2928	0.4085
Round-robin	0.3950	0.2969	0.2943	0.4129
SumRSV	**0.3980**	0.2965	0.3036	0.4134
NormMax	0.3977	0.2935	0.3010	0.4152
NormRSV (Eq.1)	0.3978	0.2937	0.3010	0.4152
Z-score	**0.3980**	0.2937	0.3014	0.4152
Z-scoreW	0.3973	0.2965	0.3009	0.4043

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

	Mean average precision			
Language	French word 49 queries	Finnish 4-gram 45 queries	Russian word 34 queries	Portuguese word 46 queries
Combination \ Translation	Comb5	Comb1	Comb4	Comb3
Okapi (#docs/#terms)	0.4071 (5/15)	0.2956 (5/30)	**0.3110** (3/15)	0.4315 (5/15)
Prosit (#docs/#terms)	0.4055 (5/10)	0.2909 (10/30)	0.2914 (3/15)	**0.4724** (10/20)
Round-robin	**0.4153**	**0.2999**	0.3007	0.4637
SumRSV	0.4096	0.2928	0.3000	0.4611
NormMax	0.4091	0.2964	0.3070	0.4711
NormRSV (Eq.1)	0.4126	0.2967	0.3086	0.4704
Z-score	0.4118	0.2955	0.3073	0.4699
Z-scoreW	0.4098	0.2948	0.3024	0.4722

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

	Russian 34 queries	Russian 34 queries	Portuguese 46 queries	Portuguese 46 queries
IR model 1 (#docs/#terms)	Prosit (3/15)	Prosit (3/15)	Prosit (10/20)	Okapi (0/0)
IR model 2 (#docs/#terms)	Okapi (3/15)	Okapi (3/10)	Okapi (5/15)	Prosit (0/0)
Data fusion operator	Round-robin	Round-robin	Norm RSV	Norm RSV
Translation tools	Free-Reverso	Pro-Free-Reverso	Onl-Free-Bab1	Onl-Free-Sys-Bab1
Mean average precision	**0.3007**	0.2962	**0.4704**	0.4491
Run name	UniNEBru1	UniNEBru2	UniNEBpt1	UniNEBpt2

Table 5: Description and mean average precision (MAP) of our official bilingual runs

2. 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 (Chen & Gey 2003) or ignoring the translation phase (Buckley *et al*. 1998), (MacNamee & Mayfield 2003); for a general overview of these questions, see (Braschler & Peters 2004). 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). After receiving a result list from each search engine, we needed to introduce a merging procedure to provide a unique ranked result list. As a first approach to this problem, we considered the round-robin approach whereby we took one document in turn from each individual list (Voorhees *et al*. 1995).

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 (Kwok *et al*. 1995). 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 into all collections, such a merging strategy may produce good retrieval performance (e.g., with a logistic regression IR model in (Chen 2003)).

Unfortunately the document scores cannot always be directly compared, thus as a third merging strategy we normalized the document scores within each collection by dividing them 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 *i*th result list, according to the following formula:

$$\text{Norm } RSV_k = ((RSV_k - \text{MinRSV}^i) / (\text{MaxRSV}^i - \text{MinRSV}^i)) \quad (1)$$

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 (Le Calvé & Savoy 2000). In our current case, we predicted the probability of relevance of 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 the S+ software), we sorted the records retrieved from separate collections in order to obtain a single ranked list. However, in order to estimate the underlying parameters, this approach requires that a training set is available. To achieve this, we used the CLEF-2003 topics and their relevance assessments in our evaluations.

$$\text{Prob}\left[D_k \text{ is rel} \mid rank_k, rsv_k\right] = \frac{e^{\beta_0+\beta_1\cdot\ln(rank_k)+\beta_2\cdot rsv_k}}{1 + e^{\beta_0+\beta_1\cdot\ln(rank_k)+\beta_2\cdot rsv_k}} \quad (2)$$

	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 model 1 (#docs/#terms)	Okapi (3/15)	Prosit (5/15)	Okapi (5/30)	Prosit (3/15)
IR model 2 (#docs/#terms)	Prosit (3/10)	Okapi (5/10)		
Data fusion operator	Z-score	Z-scoreW		
Translation tools		Bab2-Free-Rev	Bab1-Inter	Rev-Free
Mean average precision	0.5580	0.4098	0.2956	0.2914
Condition B				
IR model 1 (#docs/#terms)	Okapi (3/15)	Prosit (5/15)	Okapi (5/30)	Prosit (3/15)
IR model 2 (#docs/#terms)	Prosit (3/10)	Okapi (5/10)	Lnu-ltc (3/40)	Okapi (3/15)
Data fusion operator	Z-score	Z-scoreW	Round-robin	Round-robin
Translation tools		Bab2-Free-Rev	Bab1-Inter	Rev-Free
Mean average precision	0.5580	0.4098	0.3080	0.3007
Condition C				
IR model (#docs/#terms)	Prosit (3/10)	Prosit (5/15)	Prosit (10/30)	Prosit (3/15)
Translation tools		Bab2-Fre-Rev	Bab1-Inter	Rev-Free
Mean average precision	0.5633	0.4055	0.2909	0.2914

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

Finally, we suggest merging the retrieved documents according to the Z-score, taken from their document scores (Savoy 2003). Within this scheme, we need to compute, for the *i*th result list, the average of the RSV_k (denoted MeanRSV^i) and the standard deviation (denoted StdevRSV^i). Based on these values, we can normalize the retrieval status value of each document D_k provided by the *i*th result list, by computing the following formula:

$$\text{Z-Score } RSV_k = \alpha_i \cdot [((RSV_k - \text{MeanRSV}^i) / \text{StdevRSV}^i) + \beta^i] \text{ with } \beta^i = ((\text{MeanRSV}^i - \text{MinRSV}^i)/\text{StdevRSV}^i) \quad (3)$$

within which the value of β^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 depicts the exact parameters used to search in 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. In the top part of Table 6, it can be seen that we used a combined query translation strategy for French, Finnish and Russian languages. As described in our monolingual experiments (Savoy 2004c), we might also apply a data fusion phase before merging the result lists. Thus when searching into 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 have 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 depicts the retrieval effectiveness of various merging strategies using three different bilingual search parameter settings. In this table, the round-robin scheme will be used as a baseline. On the one hand, when different search engines are merged (Condition A and Condition B), the raw-score merging strategy results in very poor mean average precision. On the other hand, when the same search engine is used (Condition C), the resulting performance is better, but this is not the best one we should be able to achieve. The normalized score merging based on Equation 1 shows degradation over the simple round-robin approach when using parameter setting Condition B (0.1042 vs. 0.2340, or -4.9% in relative performance). Applying our logistic model using both the rank and the document score as explanatory variables, the resulting mean average precision is clearly better than the round-robin merging strategy and than other merging approaches (under Condition A or C). Under Condition B, the difference between our logistic model and the Z-score merging strategy is rather small (0.3111 vs. 0.3019, or 3.1% in relative performance).

As a simple alternative, we 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 collection (because the last three represent larger corpora). This merging strategy results in good retrieval performance, better that the simple round-robin approach. Finally, the Z-score merging approach seems to provide generally satisfactory performance. Moreover, we may multiply the Z-score by an α value (performance under the label "$\alpha_i = 1.5$" with the α_i values set as follows: EN: 1.5, FR: 1.5, FI: 1.0, and RU: 1.0).

	Mean average precision (% change)		
Parameters 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%)
Logistic 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) $\alpha_i = 1.5$	0.2669 (+11.9%)	0.3019 (+24.2%)	0.2867 (+21.6%)
Logistic reg. & Selection (0)	0.2957 (+23.9%)	0.2959 (+21.8%)	0.3405 (+44.4%)
Logistic reg. & Selection (3)	0.2953 (+23.8%)	0.2982 (+22.7%)	0.3378 (+43.3%)
Logistic reg. & Selection (10)	0.2990 (+25.3%)	0.3008 (+23.8%)	0.3381 (+43.4%)
Logistic reg. & Selection (20)	0.3010 (+26.1%)	0.3029 (+24.7%)	0.3384 (+43.5%)
Logistic reg. & Selection (50)	0.3044 (+27.6%)	0.3064 (+26.1%)	0.3388 (+43.7%)
Logistic reg. & OptimalSelect	**0.3234** (+35.5%)	**0.3261** (+34.2%)	**0.3558** (+50.9%)

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

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 cannot be found by the search engine. In a cross-lingual environment we have found 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 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 take into account these phenomena, we have designed a selection procedure which 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 submitted query and the collection). In the second step, for each result list (or language) we sum the document scores of the first 15 top-ranked documents. If this sum exceeds a given

threshold (depending on the collection or search engine), we can thus consider that the corresponding collection contains many pertinent documents. Otherwise, we might only include the *m* best ranking retrieved items from the corpus (with a relatively small *m* value). We may thus 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 "Logistic reg. & Selection (*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 "Logistic reg. & OptimalSelect", we have computed the mean average precision that can be achieved when the selection is done without any error (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 MAP 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 done based on the Prosit retrieval model in order to obtain more comparable document score across the various collections.

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

Table 8: Description and mean average precision (MAP) of our official multilingual runs

Conclusion

In this fifth CLEF evaluation campaign, we evaluated various query translation tools (see Table 1), together with a combined translation strategy (see Table 2), resulting in a retrieval performance that is worth considering. However, while a bilingual search can be viewed as easier for some language pairs (e.g., from an English query into a French document collection, or English to Portuguese), this task is clearly more complex for other language pairs (e.g., English to Finnish). Combining different result lists (see Table 3 or 4), we cannot always obtain a better retrieval effectiveness compared to isolated runs.

In multilingual tasks, searching documents written in different languages represents a real challenge. In this case we propose a new simple selecting strategy which will avoid extracting a relatively large number of documents from collections when these documents are of little interest with respect to the current request (see Table 7). In this multilingual task, it is also interesting to mention that combining the result lists provided by same search engine (Condition C in Table 7) may sometimes produce good retrieval effectiveness compared to combining different search models (Condition A in Table 7).

Acknowledgments

The authors would like to thank the CLEF-2004 task organizers for their efforts in developing various European languages test-collections. The authors 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

Braschler, M. & Peters, C. (2004). Cross-language evaluation forum: Objectives, results and achievements. *IR Journal*, 7(1-2), 7-31.

Buckley, C., Mitra, M., Waltz, J. and Cardie, C. (1998). Using clustering and superconcepts within SMART. In *Proceedings of TREC-6*, (pp. 107-124). Gaithersburg: NIST Special Publication 500-240.

Chen, A. & Gey, F. (2003). Combining query translation and document translation in cross-language retrieval. In *Proceedings CLEF-2003*, (pp. 39-48). Trondheim.

Chen, A. (2003). Cross-language retrieval experiments at CLEF 2002. In C. Peters, M. Braschler, J. Gonzalo, & M. Kluck, (Eds), *Advances in Cross-Language Information Retrieval*, (pp. 28-48), Springer-Verlag, Berlin, LNCS #2785.

Kishida, K., Kuriyama, K., Kando, N. & Eguchi, K. 2004. Prediction of performance on cross-lingual information retrieval by regression models. In *Proceedings NTCIR-4*, (pp. 219-224). Tokyo: NII.

Kwok, K.L., Grunfeld, L. & Lewis, D.D. (1995). TREC-3 ad-hoc, routing retrieval and thresholding experiments using PIRCS. In *Proceedings of TREC'3*, (pp. 247-255). Gaithersburg: NIST Publication #500-225.

Le Calvé, A. & Savoy, J. (2000). Database merging strategy based on logistic regression. *Information Processing & Management*, 36(3), 341-359.

MacNamee, P. & Mayfield, J. (2003). JHU/APL experiments in tokenization and non-word translation. In *Proceedings CLEF-2003*, (pp. 19-28). Trondheim.

Nie, J. Y., Simard, M., Isabelle, P. & Durand, R. (1999). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proceedings of the ACM-SIGIR'99*, (pp. 74-81). New York: The ACM Press.

Savoy J. (2003). Report on CLEF-2003 multilingual tracks. In *Proceedings CLEF-2003*, (pp. 7-12). Trondheim.

Savoy, J. (2004a). Combining multiple strategies for effective monolingual and cross-lingual retrieval. *IR Journal*, 7(1-2), 121-148.

Savoy, J. (2004b). Report on CLIR task for the NTCIR-4 evaluation campaign. In *Proceedings NTCIR-4*, (pp. 178-185). Tokyo: NII.

Savoy, J. (2004c). Report on CLEF-2004 monolingual tracks. In *Proceedings CLEF-2004* (this volume). Bath.

Voorhees, E.M., Gupta, N.K. & Johnson-Laird, B. (1995). The collection fusion problem. In *Proceedings of TREC'3,* (pp. 95-104). Gaithersburg: NIST Publication #500-225.

C201	Domestic Fires	C226	Sex-change Operations
C202	Nick Leeson's Arrest	C227	Altai Ice Maiden
C203	East Timor Guerrillas	C228	Prehistorical Art
C204	Victims of Avalanches	C229	Dam Building
C205	Tamil Suicide Attacks	C230	Atlantis-Mir Docking
C206	G7 Summit in Halifax	C231	New Portuguese Prime Minister
C207	Fireworks Injuries	C232	Pension Schemes in Europe
C208	"Sophie's World"	C233	Greenhouse Effect
C209	Tour de France Winner	C234	Deaf and Society
C210	Nobel Peace Prize Candidates	C235	Seal-hunting
C211	Peru-Ecuador Border Conflict	C236	A typhoon in the Philippines
C212	Sportswomen and Doping	C237	Panchen Lama
C213	Papal Travels	C238	Lady Diana
C214	Multi-billionaires	C239	Mental Health of the Young
C215	Re-election of Peru's President	C240	Sioux Ghost Shirt
C216	Glue-sniffing Youngsters	C241	New political parties
C217	AIDS in Africa	C242	Record Permanence in Space
C218	Andreotti and the Mafia	C243	Films of Kieslowski
C219	EU Commissioner Candidates	C244	Footballer of the Year 1994
C220	European Cars in Russia	C245	Christopher Reeve
C221	2002 Olympic Winter Games	C246	Castro visits UN
C222	Presidential elections in France	C247	Alexander the Great's Tomb
C223	Chernobyl Disaster outside ex-USSR	C248	Macedonia Name Dispute
C224	Woman solos Everest	C249	Women's Ten Thousand Metres Champion
C225	Nuclear Power Plant of Sosnovyi Bor	C250	Rabies in Humans

Title of the queries of the CLEF-2004 test-collection