Report on CLEF-2003 Multilingual Tracks

Jacques Savoy

Institut interfacultaire d'informatique, Université de Neuchâtel,
Pierre-à-Mazel 7, 2001 Neuchâtel, Switzerland
Jacques.Savoy@unine.ch http://www.unine.ch/info/clef/

Abstract. For our third participation in the CLEF evaluation campaign, our objective for both multilingual tracks is to propose a new merging strategy that does not require a training sample to access the multilingual collection. As a second objective, we want to verify whether our combined query translation approach would work well with new requests.

1 Introduction

Based on our experiments of last year [5], we are participating in both the small and large multilingual tracks. In the former, we retrieve documents written in the English, French, Spanish, and German languages based on a request written in one given language. Within the large multilingual track, we also had to consider documents written in Italian, Dutch, Swedish, and Finnish. As explained in Section 2, and for both multilingual tracks, we adopt a combined query translation strategy that is able to produce queries in seven European languages based on an original request written in English. After this translation phase, we search in the corresponding document collection using our retrieval scheme (bilingual retrieval) [5], [6]. In Section 3, we carry out a multilingual information retrieval, investigating various merging strategies based on the results obtained during our bilingual searches.

2 Bilingual Information Retrieval

In our experiments, we have chosen the English as the query language from which requests are to be automatically translated into seven different languages, using five different machine translation (MT) systems and one bilingual dictionary. The following freely available translation tools were used:

1. SYSTRAN™ babel.altavista.com/translate.dyn,
2. GOOGLE™ www.google.com/language_tools,
3. FREETRANSLATION™ www.freetranslation.com,
4. INTERTRAN™ www.tranexp.com:2000/InterTran,
5. REVERSO ONLINE™ translation2.paralink.com,
6. BABYLON™ www.babylon.com.

When translating an English request word-by-word using the Babylon bilingual dictionary, 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 translation tools, using the Okapi probabilistic model and based on word-based indexing scheme. 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 see that usually the `Reverso` or the `FreeTranslation` system produce interesting retrieval performance. We found only two translation tools for the Swedish and the Finnish languages but unfortunately their overall performance levels were not very good.

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

	Mean average precision						
Language	French 52 que.	German 56 que.	Spanish 57 que.	Italian 51 que.	Dutch 56 que.	Swedish 54 que.	Finnish 45 que.
Manual	51.64	44.54	48.85	48.80	46.86	40.54	46.54
Systran	40.55	32.86	36.88	35.43	N/A	N/A	N/A
Google	40.67	30.05	36.78	35.42	N/A	N/A	N/A
FreeTrans	**42.70**	31.65	39.37	**37.77**	**29.59**	N/A	N/A
InterTran	33.65	24.51	28.36	33.84	22.04	23.08	9.72
Reverso	42.55	**35.01**	**41.79**	N/A	N/A	N/A	N/A
Babylon 1	41.99	31.62	33.35	33.72	28.81	**26.89**	**9.74**
Babylon 2	39.88	31.67	31.20	27.59	27.19	20.66	N/A
Babylon 3	36.66	30.19	29.98	26.32	24.93	21.67	N/A

A particular translation tool may however produce acceptable translations for a given set of requests, but may perform poorly for other queries. This is a known phenomenon [7], even for manual translations. When studying various (manual) translations of the Bible, D. Knuth noted:

> "Well, my first surprise was that there is a tremendous variability between the different translations. I was expecting the translations do differ here and there, but I thought that the essential meaning and syntax of the original language would come through rather directly into English. On the contrary, I almost never found a close match between one translation and another. ... The other thing that I noticed, almost immediately when I had only looked at a few of the 3:16s, was that no translation was consistently the best. Each translation I looked at seemed to have its good moments and its bad moments." [2]

To date we have not been able to detect when a given translation will produce satisfactory retrieval performance and when it will fail. Thus before carrying out the retrieval, we have chosen to generate a translated query by concatenating two or more translations. Table 2 shows the retrieval effectiveness for such combinations, using the Okapi probabilistic model (word-based indexing). 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. The resulting retrieval performance is better than the best single translation scheme indicated in the row labeled "Best" (except for the strategy "Comb 1" in Spanish).

Table 2. Mean average precision of various combined translation devices (TD queries, word-based indexing, Okapi model)

	Mean average precision						
Language	French 52 que.	German 56 que.	Spanish 57 que.	Italian 51 que.	Dutch 56 que.	Swedish 54 que.	Finnish 45 que.
Comb 1	Rev+Ba1	Rev+Ba1	Rev+Ba1	Fre+Ba1	Int+Ba1	Int+Ba1	Int+Ba1
Comb 2	Rev+Sy +Ba1	Rev+Sy +Ba1	Rev+Sy +Ba1	Fre+Go +Ba1	Fre+Ba1	Int+Ba2	
Comb 2b	Rev+Go +Ba1	Rev+Go +Ba1	Rev+Go +Ba1	Fre+Int +Ba1	Fre+Ba2		
Comb 3	Rev+Go+ Fre+Ba1	Rev+Sys +Int+Ba1	Rev+Go+ Fre+Ba1	Fre+Go+ Int+Ba1	Fre+Int +Ba1		
Comb 3b	Rev+Go+ Int+Ba1	Rev+Go+ Int+Ba1	Go+Fre+ Sys+Ba2	Fre+Go+ Sys+Ba1	Fre+Int +Ba2		
Comb 3c		Rev+Sys +Fre+Ba1	Rev+Fre +Ba1				
Best	42.70	35.01	41.79	37.77	29.59	26.89	9.74
Comb 1	45.68	37.91	40.77	41.28	31.97	**28.85**	**13.32**
Comb 2	45.20	39.98	42.75	41.10	33.73	26.25	
Comb 2b	45.22	39.74	42.71	41.21	31.19		
Comb 3	**46.33**	39.25	**43.15**	**42.09**	**35.58**		
Comb 3b	45.65	39.02	42.15	40.43	34.45		
Comb 3c		**40.66**	42.72				

As described in [6], for each language, we used a data fusion search strategy using both the Okapi and Prosit probabilistic models (word-based for French, Spanish and Italian; word-based, decompounding, and n-grams for German, Dutch, Swedish and Finnish). The data shown in Table 3 indicates that our data fusion approaches usually show better retrieval effectiveness (except for the Spanish and Italian language) than do the best single IR models used in these combined approaches (row labeled "Single IR"). 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) results in relatively good retrieval performance, usually better than the mean average precision depicted in Table 3, except for the Finnish language.

Table 3. Mean average precision of automatically translated queries using various data fusion approaches (Okapi & Prosit models)

	Mean average precision						
Language	French	German	Spanish	Italian	Dutch	Swedish	Finnish
	52 que.	56 que.	57 que.	51 que.	56 que.	54 que.	45 que.
data fusion on	2 IR	3 IR	2 IR	2 IR	6 IR	6 IR	3 IR
Q combination	Comb 3b	Comb 3b	Comb 2	Comb 3	Comb 3b	Comb 1	Comb 1
Single IR	45.65	39.02	42.75	42.09	34.45	28.85	13.32
combSUM	46.37	43.02	42.09	41.18	**34.84**	**34.96**	**20.95**
combRSV%	46.29	42.68	41.96	40.50	35.51	32.04	17.74
NormN, Eq. 1	**46.30**	**43.06**	41.94	40.52	35.48	32.56	17.93
round-robin	45.94	40.41	42.18	41.42	31.89	29.88	19.35

Table 4. Mean average precision using various data fusion approaches and blind query expansion (Okapi & Prosit models)

	Mean average precision						
Language	French	German	Spanish	Italian	Dutch	Swedish	Finnish
	52 que.	56 que.	57 que.	51 que.	56 que.	54 que.	45 que.
data fusion on	2 IR	3 IR	2 IR	2 IR	6 IR	6 IR	3 IR
Q combination	Comb 3b	Comb 3b	Comb 2	Comb 3	Comb 3b	Comb 1	Comb 1
Single IR	45.65	39.02	42.75	42.09	34.45	28.85	13.32
combSUM	47.82	51.33	47.14	48.58	**43.00**	**42.93**	**19.19**
combRSV%	49.05	51.50	48.43	48.57	41.19	40.73	17.07
NormN, Eq. 1	**49.13**	**51.83**	**48.68**	**48.62**	41.32	41.53	17.21
round-robin	48.94	46.98	48.14	48.62	36.64	37.18	16.97

3 Multilingual Information Retrieval

Using the original and the translated queries, we then search for pertinent items within each of the four and eight corpora respectively. From each of these result lists and using a merging strategy, we need to produce a unique ranked result list showing the retrieved items. As a first approach, we considered the round-robin (RR) approach whereby we took one document in turn from all individual lists [8].

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 [3]. Such a strategy is called raw-score merging and produces a final list sorted by the document score computed by each collection.

Unfortunately the document scores cannot 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 "NormN"), we may normalize the document RSV_k scores within the ith result list, according to the following formula:

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

As a fifth merging strategy, we might use the logistic regression [1] to predict the probability of a binary outcome variable, according to a set of explanatory variables [4]. In our current case, we predict 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 [9]), we sort 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 be developed. To do so in our evaluations we used the CLEF-2002 topics and their relevance assessments.

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

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

$$NormZ\ RSV_k\ =\ \alpha_i\ \cdot\ \left[\frac{RSV_k\ -\ MeanRSV^i}{StdevRSV^i}\ +\ \delta_i\right] \tag{3}$$

$$\text{with } \delta_i = \frac{MeanRSV^i - MinRSV^i}{StdevRSV^i}$$

within which 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.

The justification for such a scheme is as follows. If the RSV_k distribution is linear, as shown in Table 5 and in Figure 1, there is no great difference between a merging approach based on Equation 1 or the proposed Z-score merging strategy. It is our point of view (and this point must still be verified), that such a distribution may appear when the retrieval scheme cannot detect any relevant items. However, after viewing different result lists provided from various queries and corpora, it seems that the top-ranked retrieved items usually provide a much greater RSV values than do the others (see Table 6 and Figure 2). Thus, our underlying idea is to emphasis this difference between these first retrieved documents and the rest of the retrieved items, by assigning a greater normalized RSV value to these top-ranked documents.

Table 5. Result list #1

Rank	RSV	NormZ	NormN
1	4	3.13049517	1.0
2	3.75	2.90688837	0.92857143
3	3.5	2.68328157	0.85714286
4	3.25	2.45967478	0.78571429
5	3	2.23606798	0.71428571
6	2.75	2.01246118	0.64285714
7	2.5	1.78885438	0.57142857
8	2.25	1.56524758	0.5
9	2	1.34164079	0.42857143
10	1.75	1.11803399	0.35714286
11	1.5	0.89442719	0.28571429
12	1.25	0.67082039	0.21428571
13	1	0.4472136	0.14285714
14	0.75	0.2236068	0.07142857
15	0.5	0	0

Table 7 depicts the mean average precision achieved by each single collection (or language) whether the queries used are manually translated (row labeled "Manual") or translated using our automatic translation scheme (row labeled "Auto.").

Table 8 depicts the retrieval effectiveness of various merging strategies. This data illustrates that the round-robin (RR) scheme presents an interesting performance and this strategy will be used as a baseline. On the other hand, the raw-score merging strategy results in very poor mean average precision. The normalized score merging based on Equation 1 (NormN) shows degradation over the

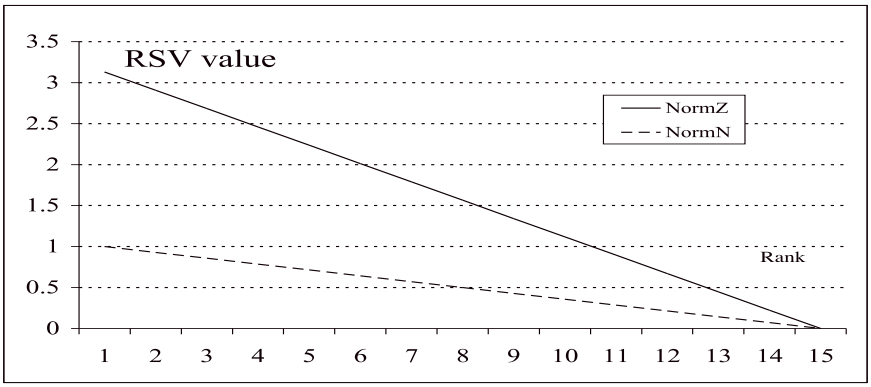

Fig. 1. Graph of normalized RSV (Result list #1)

Table 6. Result list #2

Rank	RSV	NormZ	NormN
1	10	2.57352157	1.0
2	9.9	2.54726114	0.98979592
3	9.8	2.52100072	0.97959184
4	9	2.31091733	0.89795918
5	8.2	2.10083393	0.81632653
6	7	1.78570884	0.69387755
7	6.2	1.57562545	0.6122449
8	4.5	1.12919824	0.43877551
9	3	0.73529188	0.28571429
10	2.1	0.49894806	0.19387755
11	1.4	0.31512509	0.12244898
12	1.2	0.26260424	0.10204082
13	1	0.21008339	0.08163265
14	0.5	0.07878127	0.03061224
15	0.2	0	0

Table 7. Mean average precision of each individual result lists used in our multilingual search

	Mean average precision							
Lang.	English 54 que.	French 52 que.	German 56 que.	Spanish 57 que.	Italian 51 que.	Dutch 56 que.	Swedish 54 que.	Finnish 45 que.
Manual	53.25	52.61	56.03	53.69	51.56	50.24	48.77	54.51
Auto.	53.60	49.13	51.33	48.14	48.58	43.00	42.93	19.19

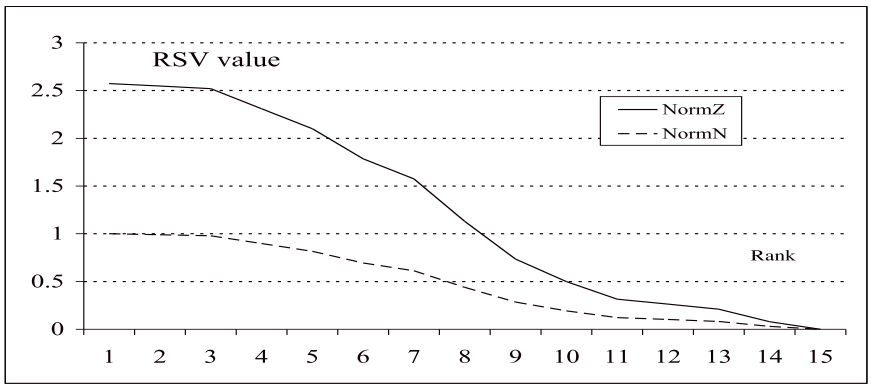

Fig. 2. Graph of normalized RSV (Result list #2)

Table 8. Mean average precision of various merging strategies

	Mean average precision (% change)			
Task	Multi-4		Multi-8	
Merging	EN, FR, DE, SP Small, manual 60 queries	EN, FR, DE, SP Small, auto. 60 queries	+IT, NL, SV, FI Large, manual 60 queries	+IT, NL, SV, FI Large, auto. 60 queries
RR, baseline	38.80	36.71	34.18	29.81
Raw-score	6.48 (-83.3%)	16.48 (-55.1%)	11.69 (-65.8%)	13.65 (-54.2%)
Norm Max	16.82 (-56.6%)	33.91 (-7.6%)	16.11 (-52.9%)	25.62 (-14.1%)
NormN (Eq. 1)	16.90 (-56.4%)	34.92 (-4.9%)	15.96 (-53.3%)	26.52 (-11.0%)
Logistic		37.58 (+2.4%)		**32.85** (+10.2%)
Biased RR	42.28 (+9.0%)	**39.20** (+6.8%)	37.24 (+9.0%)	32.26 (+8.2%)
NormZ (Eq 3)	39.44 (+1.6%)	35.07 (-4.5%)	33.40 (-2.3%)	27.43 (-8.0%)
NormZ $\alpha_i = 1.25$	41.94 (+8.1%)	37.46 (+2.0%)	36.80 (+7.7%)	29.72 (-0.3%)
NormZ $\alpha_i = 1.5$	**42.35** (+9.1%)	37.67 (+2.6%)	**37.67** (+10.2%)	29.94 (+0.4%)
NormZ coll-d	41.28 (+6.4%)	37.24 (+1.4%)	36.25 (+6.1%)	29.62 (-0.6%)

Table 9. Mean average precision of various data fusion operators on two or three merging strategies

	Mean average precision (% change)			
Task	Multi-4		Multi-8	
Data fusion	EN, FR, DE, SP Small, manual bRR, Z-1.5	EN, FR, DE, SP Small, auto. bRR, log., Z-1.5	+IT, NL, SV, FI Large, manual bRR, Z-1.5	+IT, NL, SV, FI Large, auto. bRR, log., Z.15
combSUM	42.86	38.52	37.47	31.37
combRSV%	**43.49**	38.71	**38.37**	32.65
NormN	43.45	38.68	38.36	32.55
round-robin	43.35	**40.32**	38.36	**33.68**

Table 10. Description and mean average precision (MAP) of our official runs (small multilingual runs in the top part, and large multilingual in the bottom)

Run name	Query Lang.	Form	Type	Merging	Parameters	MAP
UniNEms	English	TD	manual	biased RR		42.28
UniNEms1	English	TD	automatic	Logistic		37.58
UniNEms2	English	TD	automatic	NormZ	$\alpha_i = 1.25$	37.46
UniNEms3	English	TD	automatic	NormZ	coll-d	37.24
UniNEms4	English	TD	automatic	biased RR		**39.20**
UniNEml	English	TD	manual	biased RR		37.24
UniNEml1	English	TD	automatic	Logistic		**32.85**
UniNEml2	English	TD	automatic	NormZ	$\alpha_i = 1.25$	29.72
UniNEml3	English	TD	automatic	NormZ	coll-d	29.62
UniNEml4	English	TD	automatic	biased RR		32.26

simple round-robin approach (34.92 vs. 36.71, -4.9% in the small, automatic experiment, and 26.52 vs. 29.81, -11% in the large automatic experiment). Using our logistic model with both the rank and the document score as explanatory variables (row labeled "Logistic"), the resulting mean average precision is better than the round-robin merging strategy.

As a simple alternative, we also suggest a biased round-robin ("Biased RR" or "bRR") approach which extracts not one document per collection per round but one document for the French, English, Italian, Swedish and Finnish corpus and two from the German, Spanish and Dutch collection (representing larger corpora). This merging strategy results in interesting retrieval performance. Finally, the new Z-score merging approach seems to provide generally satisfactory performance. Moreover, we may multiply the normalized Z-score by an α value (performance under the label "NormZ $\alpha_i = 1.25$" or "NormZ $\alpha_i = 1.5$"). Under the label "NormZ coll-d", the α values are collection-dependant and are fixed as follows: EN: 1, FR: 0.9, DE: 1.2, SP: 1.25, IT: 0.9, NL: 1.15, SV: 0.95, and FI: 0.9.

Of course, we may combine the two or three best merging strategies (performance depicted in Table 8, namely the "biased round-robin" (denoted "bRR"), "logistic regression" (or "log.") and the "NormZ $\alpha_i = 1.5$" (or "Z-1.5")). Using various data fusion operators, the retrieval effectiveness of these data fusion approaches are shown in Table 9. Finally, the descriptions of our official runs for the small and large multilingual tracks are shown in Table 10.

4 Conclusion

In this fourth CLEF evaluation campaign, we have evaluated various query translation tools, together with a combined translation strategy, resulting in a retrieval performance that is worth considering. However, while a bilingual search can be viewed as easier for some pairs of languages (e.g., from an English query into

a French document collection), this task is clearly more complex for other languages pairs (e.g., English to Finnish). On the other hand, the multilingual, and more precisely the large multilingual task, shows how searching documents written in eight different languages can represent a challenge. In this case, we have proposed a new simple merging strategy based on the Z-score computed from the document scores, a merging scheme that seems to result in interesting performance.

Acknowledgments. The author would like to thank C. Buckley from SabIR for giving us the opportunity to use the SMART system. This research was supported in part by the Swiss National Science Foundation (grant #21-66 742.01).

References

1. Hosmer, D.W., Lemeshow, S.: Applied Logistic Regression. 2nd edn. John Wiley, New York (2000)
2. Knuth, D. E.: Things a Computer Scientist Rarely Talks About. CSLI Publications, Stanford (2001)
3. Kwok, K. L., Grunfeld, L., Lewis, D. D.: TREC-3 Ad-hoc, Routing Retrieval and Thresholding Experiments using PIRCS. In Proceedings of TREC'3, NIST Publication #500-225, Gaithersburg (1995) 247–255
4. Le Calvé, A., Savoy, J.: Database Merging Strategy based on Logistic Regression. Information Processing & Management, **36** (2000) 341–359
5. Savoy, J.: Report on CLEF-2002 Experiments: Combining Multiple Sources of Evidence. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.): Cross-Language Information Retrieval and Evaluation. Lecture Notes in Computer Science. Springer-Verlag, Berlin Heidelberg New York (2003) to appear
6. Savoy, J.: Report on CLEF-2003 Monolingual Tracks: Fusion of Probabilistic models for Effective Monolingual Retrieval. In *this volume*
7. Savoy, J.: Combining Multiple Strategies for Effective Cross-Language Retrieval. IR Journal, (2003) to appear
8. Voorhees, E. M., Gupta, N. K., Johnson-Laird, B.: The Collection Fusion Problem. In Proceedings of TREC'3, NIST, Publication #500-225, Gaithersburg (1995) 95–104
9. Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S-PLUS. Springer, New York (1999)