Report on CLEF 2003 Experiments at UB

Miguel E. Ruiz
State University of New York at Buffalo, School of Informatics
meruiz@buffalo.edu

Abstract

This paper presents the results obtained by the University at Buffalo (UB) in CLEF 2003. Our efforts concentrated in the monolingual retrieval and large multilingual retrieval tasks. We used a modified version of the SMART system, a heuristic bigram generation program that works across multiple languages, and pseudo relevance feedback. Query translation was performed using a publicly available machine translation software. Our results show small improvements in performance due to the use of bigrams and pseudo relevance feedback.

1 Introduction

This paper describes our experiments for the CLEF 2003 participation of the University at Buffalo. This is the first time that we participate in this conference and most of our efforts concentrated on developing the system and resources necessary for participation. We used a modified version of the SMART system [4] as a retrieval engine. This system was modified to add ISO-Latin 1 encoding as proposed by Douglas Oard [1]. We also incorporated Porter's stemmers [2] for 11 languages which are publicly available from snowball.tartarus.org.

We participated in the large multilingual retrieval tasks EN⟶X and ES⟶X and in monolingual retrieval for 8 languages Dutch, English, Finnish, French, German, Italian, Spanish and Swedish.

Section 2 describes the document processing and representation used in our system. Section 3 presents the approach used in our monolingual runs, section 4 describes the approach used for our multilingual runs. Section 5 presents our results and initial analysis. Conclusions and future work is presented in section 6.

2 Document Processing and Representation

We followed a standard processing method removing stop words, and stemming the remaining words using Porter's stemmer. Additionally we used a heuristic method to try to capture phrases and proper nouns. For this purpose we preprocessed the documents to identify fragments delimited by punctuation symbols and extract bigrams (groups of two consecutive words) that don't include stopwords. The process takes into account exceptions that allow a limited number of stop words to be part of the bigrams, i.e. "Corea del Norte". A short list of exceptions was defined for all 8 languages and consists of 18 exceptions that are presented in Table 1. An example of a query with the added bigrams is presented in Table 2.

We build two vectors (one for original words and one for bigrams) to represent the two vocabularies of each document or query. Similarity is then computed as the linear combination of the vector similarities of each vocabulary as follows:

$$Sim(\mathbf{d},\mathbf{q}) = \lambda \times Sim_{Terms}(\mathbf{d},\mathbf{q}) + \eta \times Sim_{bigrams}(\mathbf{d},\mathbf{q}) \tag{1}$$

English	”of”, ”for”
Dutch	”voor”, ”van”
French	”pour”,”de”,”d’”
Spanish	”de”, ”para”,”por”,”del”
German	”als”, ”über”, ”von”
Italian	”di”, ”il”
Finnish	”ajaksi”
Nowegian and Swedish	”av”, ”för”, ”te”

Table 1: List of stopwords allowed to be part of bigrams

<top> <num> C196 </num> <EN-title> Merger of Japanese Banks </EN-title> <EN-desc> Find reports on the merger of the Japanese banks Mitsubishi and Bank of Tokyo into the largest bank in the world. </EN-desc> <bigrams> Merger_of_Japanese Japanese_Banks Find_reports Japanese_banks banks_Mitsubishi Bank_of_Tokyo largest_bank </bigrams> </top>
<top> <num> C196 </num> <ES-title> Fusión de bancos japoneses </ES-title> <ES-desc> Encontrar documentos sobre la fusión del banco japonés Mitsubishi y el Banco de Tokyo para formar el mayor banco del mundo. </ES-desc> <bigrams> Fusión_de_bancos bancos_japoneses Encontrar_documentos fusión_del_banco banco_japonés japonés_Mitsubishi Banco_de_Tokyo Tokyo_para_formar banco_del_mundo </bigrams> </top>

Table 2: Examples of bigrams generated for a query

Performance of Short queries (T)			
Language	Terms	Terms & bigrams	Ret. Feedback
Dutch	0.3659	0.3683	0.3885
English	0.3675	0.3802	0.4035
Finnish	0.2750	0.2725	0.3095
French	0.2963	0.2919	0.3156
German	0.3124	0.3178	0.3536
Italian	0.2873	0.2800	0.3155
Spanish	0.3841	0.3833	0.4235

Performance of Medium-size queries (TD)			
Language	Terms	Terms & bigrams	Ret. Feedback
Dutch	0.4253	0.4229	0.4338
English	0.4710	0.4917	0.5142
Finnish	0.2938	0.2903	0.3116
French	0.3937	0.3931	0.4245
German	0.3637	0.3706	0.4211
Italian	0.3674	0.3589	0.4054
Spanish	0.4766	0.4799	0.5175

Table 3: Monolingual performance on CLEF2002 data

where λ and η are coefficients that weight the contribution of each vocabulary, **d** is the document vector and **q** is the query vector. We used the pivoted length normalization (Lnu.ltu) weighting scheme [6]. For our runs the *slope* was fixed to 0.25 and the *pivot* was set to the average length of the documents in the collection.

3 Monolingual Retrieval

Monolingual retrieval is crucial for any successful cross-language retrieval. For this reason we decided to explore methods that could be applied across all the supported languages. For our monolingual retrieval we use publicly available stop word lists and review them for several languages (English, Spanish, French, and Italian). We also used the publicly available versions of Porter's stemmer, which is available for 11 languages at snowball.tartarus.org, adding them to the SMART system. We tested the performance of the system on each of the 8 languages of interest in the multilingual-8 task. Results of the baselines for CLEF2002 data are presented in Table 3.

For retrieval feedback we assumed that the top 5 documents were relevant and the bottom 100 documents from the 1000 retrieved were assumed to be not relevant. We expanded the query with 30 terms of each vocabulary (Terms and Bigrams) using Rocchio's formula to rank the expansion term ($\alpha = 8$, $\beta = 64$, $\gamma = 16$, relative weight of Terms and Phrases was 5:1). These settings were determined empirically (using CLEF2002 data) across all monolingual collections.

We observed that performance was significantly improved in English by adding Bigrams. we also noticed that retrieval feedback for short and medium queries works better for most languages when we use the terms and bigrams. The only exception was Dutch in which we noticed that retrieval feedback on using only terms worked slightly better than our runs with terms and bigrams.

4 Cross-Language Retrieval

We concentrated on exploring cross-language retrieval using as base languages English and Spanish. We used a publicly available site (www.intertran.com) to translate the queries from English or Spanish to each of the other 7 languages. These queries were then preprocessed to add bigrams and the expanded query was indexed in each of the corresponding languages. For each query, the results were re scored using the following normalized score function [3, 5]:

$$rsv'_j = \frac{rsv_j - rsv_{min}}{rsv_{max} - rsv_{min}} \tag{2}$$

where rsv_j is the original retrieval score, and rsv_{max} and rsv_{min} are the maximum and minimum document score of the documents retrieved for the current query. The merging program combined all 8 files, ranked the documents by the scaled scores and selected the top 1000 ranked documents for each query.

5 Results

Official Monolingual Runs						
run name	# of queries	Avg-P	Best	Above Mean	Below Mean	Worst
UBmonoNLrf1	56	0.4180	3	26	26	1
UBmonoNLrf2	56	0.4225	3	22	27	4
UBmonoENrf1	54	0.4746	12	23	19	0
UBmonoENrf2	54	0.4488	9	21	22	2
UBmonoFIrf1	45	0.4901	5	17	21	2
UBmonoFIrf2	45	0.4790	3	17	21	4
UBmonoFRrf1	52	0.4645	7	23	22	0
UBmonoFRrf2	52	0.4638	9	22	21	0
UBmonoDErf1	56	0.4425	1	24	30	1
UBmonoDErf2	56	0.4470	1	27	27	1
UBmonoITrf1	51	0.4857	6	25	20	0
UBmonoITrf2	51	0.4965	8	23	20	0
UBmonoSVrf1	54	0.3906	3	21	29	1
UBmonoSVrf2	54	0.3910	4	18	30	2
UBmonoESrf1	57	0.1231	1	2	44	10
UBmonoESrf2	57	0.1267	1	3	43	10
Unofficial Monolingual Runs						
UBmonoESrf1 (corrected)	57	0.4852	2	35	20	0
UBmonoESrf2 (corrected)	57	0.4943	2	35	20	0
UBESmono.T.rf1	57	0.3903	3	16	38	0
UBESmono.T.rf2	57	0.3965	2	21	34	0

Table 4: Monolingual performance on CLEF2003 data

The results results that were submitted in all monolingual tasks included bigrams and pseudo-relevance feedback as explained before. Runs that end in rf1 use 30 terms to expand each query. We also tried with a more aggressive expansion strategy that uses 300 terms (runs ending in rf2).

Our official results are presented in Tables 4 and 5. We discovered a bug in our Spanish runs that significantly affected the Spanish monolingual runs as well as all our multilingual runs. The Unofficial results show the performance of the corrected runs.

In general, our monolingual performance is acceptable and performs on or above the median system. The best monolingual performance was obtained for English, French and Italian. The corrected Spanish runs also show good monolingual performance. We also have included in this table two monolingual runs with short queries (using only the title). As expected they don't perform as well as the queries based on Title and Description. The results of using the more aggressive expansion strategy are mixed. In Dutch, German, Spanish and Italian this gives small improvements in average precision but in English, Finnish and Swedish performance decreases.

In terms of multilingual performance, the corrected unofficial runs show that EN⟶X and ES⟶X have about the same average precision (0.19). Short queries, as expected, perform significantly below our standard queries. This seems to be a consequence of translation problems because of lack of context. These multilingual runs also show a third type of queries that combines the results from rf1 and rf2 queries (with 30 and 300 expansion terms respectively). The results of this strategy are mixed since it improved results for ES⟶X but reduced performance of EN⟶X queries.

6 Conclusions and Future Work

Since this was our first time participating in CLEF we have learned many lessons from our experiments. We still are working on the query by query analysis which should show us a better picture of the things that should be improved in our system. We plan to research the use of different alternatives for translation (instead of relying on a single MT system) since this seems to play an important role multilingual performance. We also need to fine tune our merging algorithm for the multilingual results.

Official Multilingual Runs						
English ⟶ X						
run name	# of queries	Avg-P	Best	Above Mean	Below Mean	Worst
UBENmultirf1	60	0.1390	1	10	48	1
UBENmultirf2	60	0.1309	1	7	52	0
UBENmultirf3	60	0.1440	1	10	49	0
UBENmultishort2	60	0.0413	0	1	37	22
UBENmultishort3	60	0.0417	0	2	38	20
Spanish ⟶ X						
UBESmultirf2	60	0.1160	0	10	36	14
UBESmultishort2	60	0.1155	0	8	44	8
Unofficial Multilingual Runs						
English ⟶ X						
UBENmulti.TD.rf1	60	0.1930	1	22	37	0
UBENmulti.TD.rf2	60	0.1857	1	18	41	0
UBENmulti.TD.rf3	60	0.1792	1	20	38	1
UBENmulti.T.rf1	60	0.0766	0	4	50	6
UBENmulti.T.rf2	60	0.0773	0	4	52	4
UBENmulti.T.rf3	60	0.0741	0	6	44	10
Spanish ⟶X						
UBESmulti.TD.rf1	60	0.1913	0	29	31	0
UBESmulti.TD.rf2	60	0.1936	1	27	32	0
UBESmulti.TD.rf3	60	0.2011	2	28	30	0
UBESmulti.T.rf1	60	0.0810	0	11	44	5
UBESmulti.T.rf2	60	0.0788	0	9	45	6
UBESmulti.T.rf3	60	0.0860	0	11	43	6

Table 5: Multilingual-8 performance on CLEF2003 data

References

[1] D. Oard. *Adaptive Vector Space Text Filtering for Monolingual and Cross-Language Applications.* PhD thesis, University of Maryland, 1996.

[2] M. F. Porter. An algorithm for suffix stripping. *Program*, 14:130–137, 1980.

[3] A. T. Powel, J. C. French, J. Callan, M. Connell, and C. L. Viles. The impact of database selection on distributed searching. In N. Belkin, P. Ingwersen, and M. Leong, editors, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 232–239, Athens, Greece, July 2000. ACM Press.

[4] G. Salton, editor. *The SMART Retrieval System: Experiments in Automatic Document Processing.* Prentice-Hall, Englewood Cliffs, NJ, 1983.

[5] J. Savoy. Report on CLEF-2002 experiments: Combining multiple sources of evidence. In C. Peters, editor, *Results of the CLEF 2002 Cross-Language System Evaluation Campaign: Working Notes for the CLEF 2002 Workshop*, Rome, Italy, September 2002.

[6] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 21–29, Zurich, Switzerland, August 1996. ACM, ACM Press.