REINA at CLEF 2007 Robust Task

Angel F. Zazo, Carlos G. Figuerola, and José L. Alonso Berrocal
REINA Research Group - University of Salamanca
C/ Francisco Vitoria 6-16, 37008 Salamanca, Spain
http://reina.usal.es

Abstract

This paper describes our work at CLEF 2007 Robust Task. We have participated in the monolingual (English, French and Portuguese) and the bilingual (English to French) subtask. At CLEF 2006 our research group obtained very good results applying local query expansion using windows of terms in the robust task. This year we have used the same expansion technique, but taking into account some criteria of robustness: MAP, GMAP, MMR, GS@10, P@10, number of failed topics, number of topics bellow 0.1 MAP, and number of topics with P@10=0. In bilingual retrieval experiments three machine translation programs were used to translate topics. For the target language, translations were merged before performing a monolingual retrieval. We also applied the same local expansion technique. This year the results were disappointing. We think out that the reason is the difficulty to select the best measurement for robustness. Perhaps the problem is that all measurements are average results over all topics, but the hard topics are inherently hard and must be analyze separately. This year all our runs also ends up in good ranking, both base runs and expanded ones. We think that the reason is that we used a good information retrieval system, and the expansion technique is robust because it does not deteriorate significantly the retrieval performance.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: *Indexing methods, Thesauruses*; H.3.3 [**Information Search and Retrieval**]: *Query formulation, Relevance feedback*; H.3.4 [**Systems and Software**]: *Performance evaluation*; I.2.7 [**Natural Language Processing**]: *Machine Translation*

General Terms

Measurement, Performance, Experimentation

Keywords

Robust Retrieval, Query Expansion, Term Windows, Association Thesauri, CLIR, Machine Translation

1 Introduction

Robust retrieval tries to obtain stable performance over all topics by focusing on poorly performing topics. Robust tracks were carried out at TREC 2003, 2004 and 2005 for monolingual retrieval [3, 4, 5], and at CLEF 2006, including monolingual, bilingual and multilingual retrieval [1]. This year only monolingual (English, French and Portuguese) and bilingual (English to French) subtask were carried out. Our research group has participated in all the subtasks. For a complete description of this task, please, see the CLEF 2007 Ad-hoc Track Overview, also published in this volume.

The system's robustness ensures that all topics obtain minimum effectiveness levels. In information retrieval the mean of the average precision (MAP) is used to measure systems' performance. But, poorly performing topics have little influence on MAP. At TREC, geometric average (GMAP), rather than MAP, turned out to be the most stable evaluation method for robustness [4]. The GMAP has the desired effect of emphasizing scores close to 0.0 (the poor performers) while minimizing differences between higher scores. Nevertheless, at the CLEF 2006 Workshop the submitted runs showed high correlations between MAP and GMAP, so at CLEF 2007 other criteria of robustness have been suggested: MAP, GMAP, P@10, number of failed topics, number of topics bellow 0.1 MAP, and number of topics with P@10=0. In our experiments we have also considered other two user-related measurements: the Generalized Success@10 (GS@10) [2], and the mean reciprocal rank (MRR). Both ones indicate the rank of the top retrieved relevant document.

Our main focus was monolingual retrieval. The steps followed are explained below. For bilingual retrieval experiments we used machine translation (MT) programs to translate topics into document language, and then we performed a monolingual retrieval.

2 Experiments

For the monolingual experiments we used the well-known vector space model, using the **dnu-ntc** term weighting scheme. For documents, letter *u* stands for the *pivoted document normalization*: we adjusted *pivot* to the average document length and *slope* set to 0.1 for all the collections. We decided to remove the terms present in more than 25 percent of documents. For the English and French languages we verified that stemming improve retrieval. Last year we saw that stemming does not deteriorate the retrieval performance of hard topics, so we also decided to apply stemming for the Portuguese language. For English we used the Porter stemmer, and for French and Portuguese the stemmers from the University of Neuchatel in the web page `http://www.unine.ch/info/clef/`. From the descriptions and narratives of the topics we automatically removed certain phrases such as "Find documents that ...", "Les documents pertinents relatent ..." or "Encontrar documentos sobre ...".

At CLEF 2006 Robust Task our research group obtained very good results applying local query expansion using windows of terms [6]. This year we have used the same expansion technique, but taking into account the new criteria. This technique uses co-occurrence relations in windows of terms from the first retrieved documents to build a thesaurus to expand the original query. Our interest was to use sort and long queries in our experiments, i.e., use the title field of the topics for sort queries, and title and description fields for long ones. A lot of tests were carried out to obtain the best performance using the training collections, but we found no settings that improve retrieval for all measurements. Then we decided to select the settings that improve the greatest number of measurements for both sort and long queries. For English the highest improvement achieved with this expansion technique was by using a distance value of 1, taking the first 15 retrieved documents to build the thesauri, and adding about 10 terms to the original query. For French, the highest improvement achieved was by using a distance value of 1, taking the first 20 retrieved documents, and adding 40 terms to the original query.

For Portuguese we decided to use the best combination obtained last year for the Spanish experiments, due two reasons. First, the Portuguese language is more similar to Spanish than English or French are. Second, the average number of terms per sentence in the Portuguese collection is very similar to the Spanish one. We use a distance value of 2, taking the first 10 documents, and adding 30 terms to the original query.

For the bilingual experiments the CLIR system was the same as that used in monolingual retrieval. A previous step was carried out before searching, to translate English topics into French. We used three MT programs: L&H Power Translator Pro 7.0, Systran[1] and Reverso[2]. For each topic we combined the terms of the translations in a single topic: this is another expansion process,

[1] `http://www.systransoft.com`
[2] `http://www.reverso.net`

Table 1: Results of the runs submitted at CLEF 2007 Robust Task.

		Basis	**Expansion***	**Basis**	**Expansion***	**Basis**
		t	t	td	td	tdn
English	MAP	0.3226	0.3205	0.3897	0.3855	0.3897
	GMAP	0.1190	0.1045	0.1850	0.1762	0.1850
(*)Settings	MRR	0.5602	0.5379	0.6922	0.6792	0.6922
for expansion:	GS@10	0.7613	0.7219	0.8506	0.8422	0.8506
distance=1	P@10	0.3200	0.3240	0.3620	0.3640	0.3620
docs=15	# failed	5	5	5	5	5
terms=10	# <0.1 MAP	16	20	7	8	7
	# P@10=0	16	23	10	11	10
French	MAP	0.3382	0.3481	0.3773	0.3804	0.3773
	GMAP	0.0940	0.0947	0.1289	0.1218	0.1289
(*)Settings	MRR	0.5749	0.5972	0.6564	0.6564	0.6564
for expansion:	GS@10	0.7555	0.7445	0.7940	0.7959	0.7940
distance=1	P@10	0.3710	0.3740	0.4140	0.4280	0.4140
docs=20	# failed	9	9	8	9	8
terms=40	# <0.1 MAP	18	19	12	12	12
	# P@10=0	23	24	19	18	19
Portuguese	MAP	0.3387	0.3533	0.4083	0.4121	0.4140
	GMAP	0.0825	0.0911	0.1369	0.1301	0.1287
(*)Settings	MRR	0.5711	0.5950	0.6286	0.6273	0.6419
for expansion:	GS@10	0.7307	0.7277	0.7855	0.7718	0.7787
distance=2	P@10	0.3013	0.3027	0.3320	0.3347	0.3360
docs=10	# failed	15	12	10	10	11
terms=30	# <0.1 MAP	28	29	22	26	23
	# P@10=0	36	39	29	30	30
EN → FR	MAP	0.3035	0.3278	0.3385	0.3455	0.3583
	GMAP	0.0821	0.0872	0.1005	0.0997	0.1228
(*)Settings	MRR	0.5819	0.6084	0.6219	0.6164	0.6794
for expansion:	GS@10	0.7555	0.7580	0.7833	0.7769	0.8096
distance=1	P@10	0.3242	0.3535	0.3770	0.3870	0.3830
docs=20	# failed	9	9	9	9	8
terms=40	# <0.1 MAP	16	16	15	14	11
	# P@10=0	22	20	19	18	16

although in most cases the three translations were identical. Finally, a monolingual retrieval was performed. The local query expansion using co-occurrence based thesauri built with terms windows was also applied.

For each subtask and topic language five runs were submitted for the test and training topics. The name of the run begins with "reina", follows the abbreviation of the language (EN, FR or PT for the monolingual runs, and E2F to indicate the English to French bilingual runs), follows the fields of the topics used in the run (t: title, td: title and description, tdn: title, description and narrative), follows with the letter "e" to indicate if expansion of terms was used and/or the letter "T" to indicate if the run is a test run. For example, the run "reinaENtdeT" stands for the test run submitted for the English collection using the title and descriptions fields of the topics, and applying term expansion. We send the "tdn" runs only for internal testing purposes.

3 Results

We only analyze results of our test runs, i.e., for the test topics of the robust task. Table 1 shows the results of the runs. We can see that term expansion no improves performance for all measurements.

4 Conclusions

At CLEF 2006 Robust Task our research group obtained very good results applying local query expansion using windows of terms in the robust task. This year at CLEF 2007 the results were disappointing. We think out that the reason is the difficulty to select the best measurement for robustness. Perhaps the problem is that all measurements are average results over all topics, but the hard topics are inherently hard and must be analyse separately. When a topic becomes hard depends on the document collection, the topic collection, the information retrieval system and the topic itself. Therefore general directives to improve performance of hard topics are difficult to suggest.

This year all our runs also ends up in good ranking, both base runs and expanded ones. We think that the reason is that we used a good information retrieval system, and the expansion technique is robust because it does not deteriorate significantly the retrieval performance.

References

[1] G. M. Di Nunzio, N. Ferro, T. Mandl, and C. Peters. CLEF 2006: Ad hoc track overview. *CLEF 2006, LNCS*, 4730, 2007.

[2] S. Tomlinson. Comparing the robustness of expansion techniques and retrieval measures. In A. Nardi, C. Peters, and J. Vicedo, editors, *ABSTRACTS CLEF 2006 Workshop, 20-22 September, Alicante, Spain. Results of the CLEF 2006 Cross-Language System Evaluation Campaign*, 2006.

[3] E. M. Voorhees. Overview of the TREC 2003 robust retrieval track. In *The Twelfth Text REtrieval Conference (TREC 2003)*, pages 69–77. NIST Special Publication 500-255, 2003.

[4] E. M. Voorhees. Overview of the TREC 2004 robust retrieval track. In *The Thirteen Text REtrieval Conference (TREC 2004), Gaithersburg, Maryland, November 16-19*. NIST Special Publication 500-261, 2004.

[5] E. M. Voorhees. Overview of the TREC 2005 robust retrieval track. In *The Fourteenth Text REtrieval Conference (TREC 2005), Gaithersburg, Maryland, November 15-18*. NIST, 2005.

[6] A. F. Zazo, J. L. Alonso Berrocal, and C. G. Figuerola. Local query expansion using terms windows for robust retrieval. *CLEF 2006, LNCS*, 4730:145–152, 2007.