A Penalisation-Based Ranking Approach for the Mixed Monolingual Task of WebCLEF 2006*

David Pinto[1,2], Paolo Rosso[1], and Ernesto Jiménez[3]

[1] Department of Information Systems and Computation, UPV, Spain
[2] Faculty of Computer Science, BUAP, Mexico
[3] School of Applied Computer Science, UPV, Spain
{dpinto, prosso}@dsic.upv.es, erjica@ei.upv.es

Abstract. This paper presents an approach of a cross-lingual information retrieval which uses a ranking method based on a penalisation version of the Jaccard formula. The obtained results after the submission of a set of runs to the WebCLEF 2006 have shown that this simple ranking formula may be used in a cross-lingual environment. A comparison with runs submitted by other teams ranks us in a third place by using all the topics. A fourth place is obtained with our best overall results by using only the new topic set, and a second place was got by using only the automatic topics of the new topic set. An exact comparison with the rest of the participants is in fact difficult to obtain and, therefore, we consider that further detailed analysis of the components should be done in order to determine the best components of the proposed system.

Keywords: Retrieval models, Mixed-Monolingual search process.

1 Introduction

The current commercial search engines, such as Google and Yahoo, provide only monolingual information retrieval and, therefore, forums dedicated to the analysis and evaluation of information retrieval systems in a cross-lingual environment, such as WebCLEF [1], are needed. Since its first edition in 2005, the WebCLEF concern has been to deal with the EuroGOV corpus, which consists in a crawl of European governmental sites from approximately 27 differents Internet domains [5,3]. The aim of this work is to evaluate a new similarity measure in the Mixed-Monolingual task of WebCLEF evaluation forum. The introduced formula is a variation of the Jaccard coefficient with a penalisation factor.

In the next section we will describe the way we have processed this corpus in order to obtain the index terms. Section 3 explains the model we have implemented, whereas its evaluation is presented in Section 4. Finally, a discussion of our participation and the obtained results in this evaluation forum are given.

* This work was partially supported by the MCyT TIN2006-15265-C06-04 project, as well as by the BUAP-701 PROMEP/103.5/05/1536 grant.

C. Peters et al. (Eds.): CLEF 2006, LNCS 4730, pp. 826–829, 2007.

2 Dataset Preprocessing

We have written two scripts in order to obtain the index terms of the EuroGOV corpus. The first script uses regular expressions for excluding all the information which is enclosed by the characters < and >. This script obtains very good results, but it is very slow and, therefore, we decided to used it only with three domains of the EuroGOV collection, namely Spanish (ES), French (FR), and German (DE). We have wrote another script based on the *html* syntax for obtaining all the terms considered valuable for indexing, which speeded up our indexing process. Unfortunately, major web pages do not strictly observe the *html* syntax and, therefore, we missed important information from those documents. We have implemented different methods for detecting the charset codification of those webpages that are not in UTF-8. We have observed that the charset codification detection is one the most difficult problem in the preprocessing step. Finally, we eliminated stop words for each language (except Greek) and punctuation symbols. The same process was applied to the queries.

3 The Penalisation-Based Ranking Approach

Nowadays, different information retrieval models are reported in literature [4] [2]. In this work, we have used a variation of the boolean model with ranking based in the Jaccard similarity formula. We named this variation "Jaccard with penalisation", because it punishes the ranking score taking into account the number of terms that a query Q_i really matches when it is compared with a document D_j of the collection. The formula used is presented as follows:

$$Score(Q_i, D_j) = \frac{|D_j| \cap |Q_i|}{|D_j| \cup |Q_i|} - \left(1 - \frac{|D_j| \cap |Q_i|}{|Q_i|}\right) \quad (1)$$

As can be seen, the first component of this formula is the typical Jaccard approximation. The evaluation of this formula is quite fast, and allows its implementation in real situations. The obtained results by using this approach are presented in the next section.

4 Experimental Results

We have submitted three different runs in order to experiment with the use of diacritics in the corpus preprocessing step. We have renamed all the runs in this paper (cursive), with respect to the names reported in [1] (bold face) as follows: *WithoutDiac* (**ERFinal**), *WithDiac* (**ERConDiac**), *CDWithoutDiac* (**DPSinDiac**).

Table 1 shows the results obtained with each of the three different approximations submitted. The *WithoutDiac* run eliminates all diacritics in both, the

corpus and the topics, whereas the *WithDiac* run only supresses the diacritics in the corpus. We may observe an expected reduction of the Mean Reciprocal Rank (MRR), but it does not differ significantly from the first run. This is clearly derived from the amount of diacritics introduced in the evaluation topics set, which is not very high. An analysis of the queries in real situations may be interesting in order to determine whether the topics set is realistic. The last run (*CDWithoutDiac*) eliminates diacritization in both, the topics and corpus; besides, it tries a charset detection for each indexed document.

Table 1. Evaluation of each run submitted

		Average Success at					
Team	**Run**	**1**	**5**	**10**	**20**	**50**	**MRR over 1939**
rfia	*WithoutDiac*	0,0665	0,1423	0,1769	0,2192	0,2625	**0,1021**
rfia	*WithDiac*	0,0665	0,1372	0,1717	0,2130	0,2568	0,1006
rfia	*CDWithoutDiac*	0,0665	0,1310	0,1681	0,1996	0,2470	0,0982

Table 2 shows a summary of all the best participant runs submitted to the mixed monolingual task of WebCLEF 2006. The Mean Reciprocal Rank (MRR) scores are reported for both the original and the new topic set. The first column indicates the name that each team had in the evaluation forum, whereas the second column indicates the name of their best run. The scores shown in that table rank us in a third place.

Table 2. Best runs for each WebCLEF 2006 participant in the mixed monolingual task

Team Name	**Run Name**	**MRR for the original topic set**	**MRR for the new topic set**
isla	CombPhrase	0.2001	0.3464
hummingbird	humWC06dpcD	0.1380	0.2390
rfia	WithoutDiac (ERFinal)	0.1021	0.1768
depok	UI2DTF	0.0918	0.1589
ucm	webclef-run-all-2006	0.0870	0.1505
hildesheim	UHiBase	0.0795	0.1376
buap	allpt40bi	0.0157	0.0272
reina	USAL_mix_hp	0.0139	0.0241

In Table 3(a) we can see the best overall results by using only the new topic set. Here we have obtained a fourth place, according to the average among the automatic and the manual topic scores. Whereas, in Table 3(b) we may observe the results by using only the new automatic generated topics. Our second place shows that the penalisation-based ranking is working well for the task proposed in this evaluation forum.

Table 3. Best overall runs for each WebCLEF 2006 participant by using the new topic set with: *(a)* all topics, and *(b)* only the automatically generated topics

Team Name	all	auto	manual	average	auto-uni	auto-bi
isla	0.3464	0.3145	0.4411	0.3778	0.3114	0.3176
hummingbird	0.2390	0.1396	0.5068	0.3232	0.1408	0.1384
depok	0.1589	0.0923	0.3386	0.2154	0.1024	0.0819
rfia	0.1768	0.1556	0.2431	0.1993	0.1568	0.1544
hildesheim	0.1376	0.0685	0.3299	0.1992	0.0640	0.0731
ucm	0.1505	0.1103	0.2591	0.1847	0.1128	0.1077
buap	0.0272	0.0080	0.0790	0.0435	0.0061	0.0099
reina	0.0241	0.0075	0.0689	0.0382	0.0126	0.0022
	(a)				*(b)*	

5 Discussion

We have proposed a new approach for the ranking formula in an information retrieval system which is based on the Jaccard formula, but with a penalisation factor. After evaluating this approach in the approximately 75% of queries from the WebCLEF evaluation forum, we have obtained the third place in the overall results, among eight participant teams. In the particular case of the new topic set, we obtained fourth place in overall (all topics), and second place using automatically generated topics. An evaluation of the use of diacritization in the task has shown that results are not significantly different, which may be suggesting that the set of queries provided for the evaluation does not have a high number of diacritics. Further investigation would determine whether this behaviour is realistic or must be tuned in further evaluations. Finally, we consider that the system proposed may be improved by taking into account a better understanding of the preprocessing phase.

References

1. Balog, K., Azzopardi, L., Kamps, J., de Rijke, M.: Overview of WebCLEF 2006. In this Volume (2007)
2. Kraaij, W., Simard, M., Nie, J.Y.: Embedding Web-based Statistical Translation Models in Cross-Language Information Retrieval. Computational Linguistics 29(3), 381–419 (2003)
3. Pinto, D., Jiménez-Salazar, H., Rosso, P.: BUAP-UPV TPIRS: A System for Document Indexing Reduction on WebCLEF. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
4. Salton, G.: Automatic Text Processing. Addison-Wesley, Reading (1989)
5. Sigurbjörnsson, B., Kamps, J., de Rijke, M.: EuroGOV: Engineering a Multilingual Web Corpus. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)