Dublin City University at CLEF 2005: Multi-8 Two-Years-On Merging Experiments

Adenike M. Lam-Adesina Gareth J. F. Jones
School of Computing, Dublin City University, Dublin 9, Ireland
{adenike,gjones}@computing.dcu.ie

Abstract. This year Dublin City University participated in the CLEF 2005 Mulit-8 Two-Years-On multilingual merging task. The objective of our experiments was to test a range of standard techniques for merging ranked lists of retrieved documents to see if consistent trends emerge for lists generated using different information retrieval systems. Our results show that the success of merging techniques can be dependent on the retrieval system used, and in consequence the best merging techniques to adopt cannot be recommended independent of knowing the retrieval system to be used.

1 Introduction

Multilingual information retrieval (MIR) refers to the process of retrieving relevant documents from collections in different languages in response to a user request in a single language. Standard approaches to MIR involve either translating the search topics into the document languages, performing cross-language information retrieval (CLIR), and then merging the ranked document sets produced for each language to form a single multilingual retrieved list, or translating the document collections into the expected topic language merging the translated collections, and then effectively performing monolingual information retrieval in the topic language. In CLEF 2003 we showed that translating the document collections into the topic language using a standard machine translation system and then merging them to form a single collection for retrieval, can result in better retrieval performance than translating the topics and then merging after CLIR retrieval [1]. However, document translation is not always practical, particularly if the collection is very large or the translation resources are limited. For MIR using topic translation and merging retrieved lists of potentially relevant documents, the different statistics of the individual collections and the varied topic translations mean that the scores of documents in the separate lists will generally be incompatible, and thus that merging is a non-trivial process.

The CLEF 2005 Multilingual merging task aims to encourage researchers to focus directly on the merging problem. Retrieval results for merged collections of noisy document translations illustrate the level of retrieval effectiveness that is possible for MIR tasks. Many CLIR experiments using topic translation have demonstrated high levels of effectiveness relative to monolingual information retrieval for individual languages. The challenge for merging is to reliably achieve similar or better MIR by

combining CLIR results, than using a single combined collection of translated documents.

Merging strategies explored previously for multilingual retrieval tasks at CLEF and elsewhere have generally produced disappointing results. Previously standardised evaluation tasks incorporating multilingual merging have been combined with the document retrieval stage. It has thus not been possible to distinguish quality of retrieval from the effectiveness of merging, or any dependency between the retrieval methods adopted and the most effective merging algorithm. The idea of the CLEF 2005 merging task is to explore the merging of provided precomputed ranked lists to enable direct comparison of the behaviour of merging strategies between different retrieval systems.

Many different techniques for merging separate result lists to form a single list have been proffered and tested in recent years. All of the techniques suggest that making an assumption that the distribution of relevant documents in the results sets of retrieval from individual collections is similar is not true [2]. Hence, straight merging of relevant documents from the sources will result in poor combination. However, none of the proposed more complex merging techniques have really been demonstrated to be consistently effective.

For our participation in the merging track at CLEF 2005 we applied a range of standard merging strategies to the two provided sets of ranked lists. Our aim was to compare the behaviour of these methods for the two sets of ranked documents in order to learn something about concepts that might be consistently useful or poor when merging ranked lists.

This paper is organized as follows: Section 2 overviews the merging techniques explored in this paper, Section 3 gives our experimental results, and Section 4 draws conclusions and considers strategies for further experimentation.

2 Merging Strategies

The aim of a merging strategy for MIR is to include as many relevant documents at the highest ranks in the merged list as possible. This section overviews the merging strategies used in our experiments. The basic idea is to modify the scored weight of each retrieved document to take account of the characteristics of the retrieval methods used to generate it, or the collection from which it has been retrieved to improve the compatibility of scores before combining the lists.

This score adjustment may take account of factors such as maximum and/or minimum matching scores in each list, or the distribution of matching scores in each list. Another factor available is to select documents for inclusion in the combined list in proportion to the relative size of the collections from which they are drawn. This works on the assumption that similar relative number of relevant documents will be found in each collection. While the process for search topic generation for the multilingual CLEF tasks mean that this will often be a reasonable assumption for these tasks, it will more however often not be the case for many topics in working systems. We include exploration of all these factors to explore their effectiveness for multilingual merging in CLEF tasks.

The schemes used in our experiments were as follows:

$$p = doc\_wgt$$

$$t = doc\_wgt * rank$$

$$d = \frac{doc\_wgt - \min\_wt}{\max\_wt - \min\_wt}$$

$$r = (\frac{doc\_wgt - \min\_wt}{\max\_wt - \min\_wt}) * rank$$

$$q = (\frac{doc\_wgt - g\min\_wt}{g\max\_wt - g\min\_wt}) * rank$$

$$b = \frac{doc\_wgt - \min\_wt}{\max\_wt - \min\_wt * rank}$$

$$m1 = (\frac{doc\_wgt - gmean\_wt}{gstd\_wt}) + (\frac{gmean\_wt - g\min\_wgt}{gstd\_wt})$$

$$m2 = (m1) * rank$$

doc_wgt = the initial document weight
gmax_wt = the global maximum weight, i.e. the highest document weight from all collections for a given query
gmin_wt = the global minimum weight, i.e. the lowest document weight from all collections for a given query
gmean_wt = the global median weight, i.e. the mean document weight from all collections for a given query,

$$gmean\_wt = \frac{\sum_{i=0}^{n} doc\_wgt_i}{totdocs}$$

totdocs = total number of retrieved documents per query across all retrieval methods
max_wt = the individual collection maximum weight for a given query
min_wt = the individual collection minimum weight for a given query

gstd_wt = the standard deviation weight calculated as,

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(doc\_wgt_i - gmean\_wt)^2}$$

rank = a parameter to control the effect of size of collection - a collection with more documents gets a higher rank (value ranges between 1 and 1.5).

where *p, t, d, r, q, b, m1 and m2* are the new document weight for all documents in all collections, and the results are labelled with the appropriate letter for the new document weight used.

Method *p* is used as a baseline using the raw document scores from the retrieved lists without modification. A useful merging scheme should be expected to improve on the performance of the *p* scheme. The *rank* factor was adjusted empirically using the 20 training topics provided for the merging task.

3 Experimental results

Results for our experiments using these merging schemes are shown in Tables 1 and 2. Our official submissions to CLEF 2005 are marked *.

Table 1. Merging results using the provided Hummingbird ranked lists

Run-id	P10	% chg.	P30	% chg.	MAP	% chg.	Rel. Ret.	chg.
dcu.hump*	0.518	-	0.396	-	0.2086	-	2982	-
dcu.humd	0.373	-28.0	0.347	-12.4	0.1775	-14.9	2965	-17
dcu.humr	0.455	-12.1	0.364	-8.0	0.1932	-7.4	2964	-18
dcu.humq	0.4576	-11.6	0.363	-8.2	0.2005	-3.9	2752	-230
dcu.humb	0.320	-32.2	0.293	-26.1	0.1596	-23.5	2950	-32
dcu.humt*	0.408	-21.3	0.328	-17.3	0.1734	-16.9	2442	-540
dcu.humm1*	0.480	-7.2	0.382	-3.6	0.1988	-4.7	2873	-109
dcu.humm2*	0.465	-10.1	0.363	-8.4	0.1846	-11.5	2846	-136

Table 2. Merging results using the provided Prosit ranked lists from the University of Neuchâtel

Run-id	P10	% chg.	P30	% chg.	MAP	% chg.	Rel. Ret.	chg.
dcu.Prositqgp*	0.450	-	0.446	-	0.3103	-	4404	-
dcu.Prositqgd	0.485	+7.7	0.444	-0.4	0.2931	-5.5	4552	+148
dcu.Prositqgr	0.495	+10.0	0.446	0.0	0.3011	-3.0	4544	+140
dcu.Prositqgq	0.465	+3.3	0.446	+0.1	0.3192	+2.9	4469	+65
dcu.Prositqgb	0.472	+5.0	0.441	-1.1	0.2834	-8.7	4538	+134
dcu.Prositqgt*	0.460	+2.2	0.446	0.0	0.3201	+3.2	4477	+73
dcu.Prositqgm1*	0.475	+5.6	0.459	+3.0	0.3241	+4.5	4486	+82
dcu.Prositqgm2*	0.470	+4.4	0.461	+3.4	0.3286	+5.9	4512	+108

Tables 1 and 2 show merging results using CLIR runs provided by Hummingbird and the University of Neuchâtel respectively. Results are shown for precision at cutoff of 10 and 30 documents, Mean Average Precision (MAP) and the total number of relevant documents retrieved. The raw score merging scheme *p* is taken as a baseline and changes for each scheme are shown for each data set with respect to the reported metrics.

The most obvious results are that the more complex merging schemes are shown in Table 2 to generally improve performance by a small amount for the Prosit data, but in Table 1 in all cases reduce performance for the Hummingbird data with respect to both the precision measures and the number of relevant retrieved. This appears to offer an answer to one of the questions associated with the CLEF merging task, namely whether the same merging techniques will always be found to be effective for different sets of ranked lists for a common merging task generated using alternative information retrieval systems. The reasons for this difference in behaviour need to be investigated. This analysis will hopefully provide insights into the selection of appropriate merging strategies or the development of merging strategies which will operate more consistently when merging different sets of ranked lists. There are some other observations of consistent behaviour which can be made be. It can be seen that there is no consistent relationship between the variation in precision measures and the number of relevant documents retrieved for the different merging schemes. Schemes with better precision can be accompanied by lower relevant retrieved and vice versa. This is most notable for the *b* results where good relevant retrieved (in relative terms) is accompanied by a large reduction in MAP for both data sets.

4 Conclusions

Results of our merging experiments for CLEF 2005 indicate that the behaviour of merging schemes varies for different sets of ranked lists. The reasons for this behaviour are not obvious, and further analysis is planned to attempt to better understand this behaviour as a basis for the extension of these techniques for merging or the proposal of new ones.

References

1. Di Nunzio, G.M., Ferro, N, and Jones, G.J.F.: CLEF 2005: Multilingual Track Overview, Proceedings of the CLEF 2005: Workshop on Cross-Language Information Retrieval and Evaluation, Vienna, Austria, 2005.
2. Lam-Adesina, A.M. and Jones, G.J.F.: Exeter at CLEF 2003: Experiments with Machine Translation for Monolingual, Bilingual and Multilingual Retrieval, Proceedings of the CLEF 2003 Workshop on Cross-Language Information Retrieval and Evaluation, Trondheim, Norway, pages 271-285, 2003.
3. Savoy, J.: Report on CLEF-2003 Multilingual Tracks, Proceedings of the CLEF 2003 Workshop on Cross-Language Information Retrieval and Evaluation, Trondheim, Norway, pages 64-73, 2003.