Using key phrases as new queries in building relevance judgments automatically

Mireille Makary[1], Michael Oakes[1], and Fadi Yamout[2]

[1]Research Group in Computational Linguistics, University of Wolverhampton, UK
{m.makary, Michael.oakes}@wlv.ac.uk

[2]Computer Science Department, Lebanese International University, Lebanon
fadi.yamout@liu.edu.lb

Abstract. We describe a new technique for building a relevance judgment list (qrels) for TREC test collections with no human intervention. For each TREC topic, a set of new queries is automatically generated from key phrases extracted from the top k documents retrieved from 12 different Terrier weighting models when the initial TREC topic is submitted. We assign a score to each key phrase based on its similarity to the original TREC topic. The key phrases with the highest scores become the new queries for a second search, this time using the Terrier BM25 weighting model. The union of the documents retrieved forms the automatically-build set of qrels.

Keywords: Evaluation, automatic qrels, key phrases, relevance judgments.

1 Introduction

We propose a new technique based on Efron's [1] work which used query aspects to automatically build a set of qrels. The qrels did not involve any human intervention but the query aspects created for each TREC topic were mostly created manually. To explain what an aspect is, consider TREC topic 402 that has "behavioral genetics" as its title. The same information need might be represented by different aspects such as "behavioral disorders" or "genetics addictions". Each manually derived aspect was considered as a query and the union of the top 100 documents retrieved for each topic was considered to be the set of "pseudo-qrels" or "aspect qrels". We generate these new query aspects automatically from key phrases extracted from documents and use them to generate a relevance judgment list.

2 Experiments

Following Efron, we use the TREC-8 and TREC-7 test collections. We start initially by submitting each TREC topic to 12 weighting models found in Terrier (BM25, DFR_BM25, LGD, In_expC2, In_expB2, IFB2, TFIDF, LemurTF_IDF, PL2, BB2, DLH13 and DLH) [3] as surrogates for different information retrieval systems. The top K (K=10) documents retrieved by all 12 weighting models are collected in a set

(S) because they have a high probability of being relevant to the topic. Next, we extract 25 keyphrases using KEA [2] from each document in (S) where each key phrase consists of 3-5 terms for TREC-8 and 2-3 terms for TREC-7. Values were determined empirically. We assign a score to each keyphrase depending on its similarity to the initial topic. We then select the key phrases with the highest scores for each topic (s >=0.4 for TREC-8 and s>=0.33 for TREC-7) and put them in a set Q. The key phrases in Q are submitted as queries to the BM25 weighting model and since we are using another query for the same topic, this leads to new relevant documents that were not retrieved in the initial topic submission. We combine the union of the documents retrieved by the key phrases in Q. These documents are considered to be the newly generated qrels for the initial topic. To compare with Efron, we used a subset of the TREC systems, the "automatic" runs. In TREC-8 there were 116 automatic runs and in TREC-7 there were 86. We computed the MAP values using the original qrels for the test collection and then the MAP values using the newly generated qrels. We ranked the systems and computed the correlation with the TREC rankings. As shown in table 1, for TREC-7 the newly generated qrels provide a better correlation than those generated from Efron's aspects, while for TREC-8 they are similar. This is acceptable considering that there is no human intervention in our method.

	Efron's aspects qrels		**Keyphrases generated qrels**	
Test Collection	**Kendall's tau**	**Spearman**	**Kendall's tau**	**Spearman**
TREC-7	0.867	0.974	**0.914**	**0.986**
TREC-8	**0.77**	**0.92**	0.762	0.912

Table 1: Kendall's tau for TREC-7 and TREC-8 automatic runs for different techniques

3 Conclusion

In this paper, we automatically generated aset of qrels based on keyphrases extracted from documents retrieved from 12 Terrier models for a particular topic and we used them as new queries instead of formulating new ones manually. The union of the documents obtained after this process was proven to be better than the aspect qrels generated by Efron. Future work can include testing this method on non-English and non-TREC test collections to evaluate its performance for any test collection.

4 References

1. Efron M.: Using multiple query aspects to build test collections without human relevance judgements, ECIR 2009
2. Witten I.H., Paynter G.W., Frank E., Gutwin C. and Nevill-Manning C.G. (2000) "KEA: Practical automatic keyphrase extraction." Working Paper 00/5, Department of Computer Science, the University of Waikato.
3. Ounis I., Amati G., Plachouras V., He B., Macdonald C. and Johnson D. Terrier Information Retrieval Platform. In Proceedings of the 27th European Conference on Information Retrieval (ECIR 05).