

EXETER AT CLEF 2003: Experiments with Machine Translation for Monolingual, Bilingual and Multilingual Retrieval

Adenike M. Lam-Adesina, Gareth J. F. Jones*
Department of Computer Science
University of Exeter EX4 4QF
United Kingdom
{A.M.Lam-Adesina, G.J.F.Jones}@ex.ac.uk

Abstract

The University of Exeter group participated in the monolingual, bilingual and multilingual-4 retrieval tasks this year. The main focus of our investigation this year was the small multilingual task comprising four languages, French, German, Spanish and English. We adopted a document translation strategy and tested four different merging techniques to combine results from the different sources to achieve an optimal performance. For both the monolingual and bilingual tasks we explored the use of a parallel collection for query expansion and term weighting and also experimented with updating synonym information to conflate British and American English word spellings.

1 Introduction

This paper describes our experiments for CLEF 2003. This year we participated in the monolingual, bilingual and multilingual retrieval tasks. The main focus of our participation this year was the multilingual task (being our first participation in this task), our submissions for the other two tasks build directly from our work from past experiments (CLEF 2001 and CLEF 2002). Our official submissions included monolingual runs for Italian, German, French and Spanish, bilingual German to Italian and Italian to Spanish, and the small multilingual tasks comprising English, French, German and Spanish collections.

Our general approach was to use translation of both collections and topics into a common language. Thus the document collections were translated into English using Systran Version:3.0 Machine Translator (Sys), and all topics translated into English using either Systran Version:3.0 or Globalink Power Translation Pro Version 6.4 (Pro) Machine Translator (MT) systems.

Following from our successful use of Pseudo-Relevance Feedback methods in past CLEF exercises (CLEF 2001, 2002) and supported by past research work in text retrieval exercises [1][2][3], we continued to use this method with success for improved retrieval. In our previous experimental work [4][5] we demonstrated the effectiveness of a new PRF method of term selection from document summaries, and found it to be more reliable than query expansion from full documents, this method is again used in the results reported here.

Following from last year, we again investigated the effectiveness of query expansion and term estimation from a parallel (pilot) collection [6] and found that caution needs to be exercised when using the collections to achieve improve retrieval for translated documents.

The remainder of this paper is structured as follows: in Section 2 we present our system setup and the information retrieval methods used, Section 3 describes the pilot search strategy, Section 4 presents and discusses experimental results and Section 5 concludes the paper with a discussion of our findings

2 System Setup

The basis of the experimental system was the City University research distribution version of the Okapi system. The documents and search topics were processed to remove stopwords from a list of about 260 words; suffix stripped using the Okapi implementation of Porter stemming [7] and terms were indexed using a small set of synonyms. Since the English document collection for CLEF 2003 incorporates both British and American documents, the synonym table was updated this year to include some common British words that have different American nomenclature.

* now at School of Computing, Dublin City University, Ireland email: Gareth.Jones@computing.dcu.ie

2.1 Term Weighting

Document terms are weighted using the Okapi BM25 weighting scheme developed in [8] and further elaborated in [9] and calculated as follows,

$$cw(i, j) = \frac{cfw(i) \times tf(i, j) \times (K1 + 1)}{K1 * ((1 - b) + (b \times ndl(j))) + tf(i, j)} \quad (1)$$

where $cw(i, j)$ represents the weight of term i in document j , $cfw(i)$ is the standard collection frequency weight, $tf(i, j)$ is the document term frequency, and $ndl(j)$ is the normalized document length. $ndl(j)$ is calculated as $ndl(j) = dl(j)/avdl$ where $dl(j)$ is the length of j and $avdl$ is the average document length for all documents. $k1$ and b are empirically selected tuning constants for a particular collection. $k1$ is designed to modify the degree of effect of $tf(i, j)$, while constant b modifies the effect of document length. High values of b imply that documents are long because they are verbose, while low values imply that they are long because they are multi-topic. In our experiments values of $k1$ and b are estimated based on the CLEF 2002 data.

2.2 Pseudo-Relevance Feedback

Retrieval of relevant documents is usually affected by short or imprecise queries. Relevance Feedback (RF) via query expansion, aims to improve initial query statements by addition of terms from user assessed relevant documents. These terms are assessed using document statistics and usually describe the information request better. Pseudo-Relevance Feedback (PRF) whereby relevant documents are assumed and used for query expansion is on average found to give improvement in retrieval performance although this is usually smaller than that observed for true user based RF.

The main implementation issue for PRF is the selection of appropriate expansion terms. In PRF problems can arise if assumed relevant documents are indeed non-relevant thus leading to selection of inappropriate terms. However, the selection of such documents might suggest partial relevance, thus, term selection from relevant section might prove more beneficial.

Our query expansion method selects terms from summaries of the top 5 ranked documents. The summaries were generated using the method described in [4]. The summary generation method combines the Luhn's Keyword Cluster Method [10], Title terms frequency method [4], Location/header method [11] and the Query-bias method [12] to form an overall significance score for each sentence. For all our experiments we used the top 6 ranked sentences as the summary of each document. From this summary we collected all non-stopwords and ranked them using a slightly modified version of the Robertson selection value (rsv) [13] reproduced below. The top 20 terms were then selected in all our experiments.

$$rsv(i) = r(i) \times rw(i) \quad (2)$$

where $r(i)$ = number of relevant documents containing term i

$rw(i)$ is the standard Robertson/Sparck Jones relevance weight [12] reproduced below

$$rw(i) = \log \frac{(r(i) + 0.5)(N - n(i) - R + r(i) + 0.5)}{(n(i) - r(i) + 0.5)(R - r(i) + 0.5)}$$

where $n(i)$ = the total number of documents containing term i

$r(i)$ = the total number of relevant documents term i occurs in

R = the total number of relevant documents for this query

N = the total number of documents

In our modified version, although potential expansion terms are selected from the summaries of the top 5 ranked documents, they are ranked using the top 20 ranked documents from the initial run.

3 Pilot Searching

Query expansion is aimed at improving initial search topics in order to make it a better expression of user's information need. This is normally achieved by adding terms selected from assumed relevant documents retrieved from the test collection, to the initial query. However, it has been shown [14] that if additional documents are available these can be used in a pilot set for improved selection of expansion terms. The underlying assumption in this method is that a bigger collection than the test collection can help to achieve better term expansion and/or more accurate parameter estimation, and hopefully better retrieval and document ranking. Based on this assumption we explore the idea of pilot searching in our CLEF experiments. The Okapi submissions for the TREC-7 [6] and TREC-8 [14] ad hoc tasks used the TREC disks 1-5, of which the document test set is a subset, for parameter estimation and query expansion. The method was found to be very effective. In order to explore the utility of pilot searching for our experiments, we used the TREC-7 and TREC-8 ad hoc document test collection itself for our pilot runs. The pilot searching procedure is as carried out as follows:

1. Run the unexpanded initial query on the pilot collection using BM25 without feedback
2. Extract terms from the summaries of the top R assumed relevant documents
3. Select top ranked terms using (3) based on their distribution in the pilot collection
4. Add desired number of selected terms to initial query
5. Store equivalent pilot weight of terms
6. Either apply expanded query to the test collection and estimate weight based on test collection or Apply expanded query and estimated weight from pilot collection on the test collection

4 Experimental results

This section describes the establishment of the parameters of our experimental system and gives results from our investigations for CLEF 2003 monolingual, bilingual and multilingual tasks. We report procedures for system parameters selection, baseline retrieval results for all languages and translation systems without the application of feedback. Corresponding results after the application of different methods of feedback including results for term weight estimation from pilot collections. The CLEF 2003 topics consist of three fields: Title, Description and Narrative. All our experiments use the Title and the Description fields only. For all runs we present the average precision results (Avep), the % change from results for baseline no feedback runs (% chg) and the number of relevant documents retrieved out of the total number of relevant in collection (Rel_ret).

4.1 Selection of System Parameters

To set appropriate parameters for our runs, development runs were carried out using the CLEF 2002 collections. These document collections consist of those used for CLEF 2001 runs and are the same as those used for CLEF 2002. For CLEF 2003 more documents were added to all individual collections, and thus we are assuming that these parameters are suitable for these modified collections as well. The Okapi parameters were set as follows $k1=1.4$ $b=0.6$. For all our PRF runs, 5 documents were assumed relevant for term selection and document summaries comprised the best scoring 6 sentences in each case. Where the length of sentence was less than 6, half of the total number of sentences was chosen. The rsv values to rank the potential expansion terms were estimated based on the top 20 ranked assumed relevant documents. The top 20 ranked expansion terms taken from these summaries were added to the original query in each case. Based on results from our previous experiments, the original topic terms are upweighted by a factor of 3.5 relative to terms introduced by PRF. In our test runs we experimented with updated synonym information to conflate British and American English word spellings. This method resulted in a further 4% improvement in average precision compared to the baseline no feedback results for our English monolingual unofficial run for CLEF 2002¹. We anticipate this being a useful technique for CLEF 2003 as well, and the updated synonym list is again used for all our experiments reported here.

¹ Given that the CLEF 2002 English collection contains only American English documents, we found this improvement in performance from spelling conflation a little surprising for the CLEF 2002 task, and we intend to carry our further investigation into the specific sources of the improvement in performance.

4.2 Monolingual runs

We submitted runs for four languages (German, French, Italian and Spanish) in the monolingual task. Official runs are marked with a * and additional unofficial runs are presented. In all cases, results are presented for the following:

1. Baseline run without feedback (exe*base)
2. Feedback runs using expanded query and term weights from the target collection (exe*mono)
3. Feedback runs using expanded query from pilot collection and term weights from test collection (exe*tcmono)
4. Feedback runs using expanded query and term weights from pilot collection (exe*tcqywg)
5. An additional Feedback run is presented where query is expanded using a pilot run on a merged collection of all four text collection comprising the small multilingual collections. (exe*comqy) with the terms weights being taken from the test collection.

Note: * refers to the target language e.g sp -> Spanish, de-> German, it->Italian and fr->French. Results are presented for both Sys and Pro MT systems

4.2.1 German Monolingual runs

Run-id	Sys MT			Pro MT		
	Avep	% chg	R-ret	Avep	% chg	R-ret
<i>Exedebase</i>	488	-	1706	441	-	1580
<i>Exedemono</i>	568*	+16.4%	1747	511*	+15.9%	1657
<i>Exedtcmono</i>	512*	+4.9%	1727	457	+3.6%	1616
<i>Exedtcqywg</i>	458	-6.1%	1665	431	-2.3%	1575
<i>Exedecomqy</i>	550	+12.7%	1751	494	+12.0%	1663

Table 1 Retrieval results for topic translation for German monolingual runs for both Sys and Pro MT, before and after applications of different feedback strategies.

4.2.2 French Monolingual runs

Run-id	Sys MT			Pro MT		
	Avep	% chg	R-ret	Avep	% chg	R-ret
<i>Exefrbase</i>	487	-	918	422	-	885
<i>Exefrmono</i>	521*	+6.9%	933	457*	+8.3%	897
<i>Exefrtcmono</i>	491*	+0.8%	921	403	-4.5%	890
<i>Exefrtcqywg</i>	489	+0.4%	920	426	+0.9%	885
<i>Exefrcomqy</i>	519	+6.6%	931	446	+5.7%	893

Table 2 Retrieval results for topic translation for French monolingual runs for both Sys and Pro MT, before and after applications of different feedback strategies.

4.2.3 Italian Monolingual runs

Run-id	Sys MT			Pro MT		
	Avep	% chg	R-ret	Avep	% chg	R-ret
<i>Exeitbase</i>	419	-	761	387	-	742
<i>Exeitmono</i>	494*	+17.9%	787	449*	+16.0%	759
<i>Exeittcmono</i>	432*	+3.1%	762	402	+3.89%	745
<i>Exeittcqywg</i>	393	-6.2%	754	387	0%	735
<i>Exeitcomqy</i>	456	+8.8%	771	452	+16.8%	759

Table 3 Retrieval results for topic translation for Italian monolingual runs for both Sys and Pro MT, before and after applications of different feedback strategies.

4.2.4 Spanish Monolingual runs

Run-id	Sys MT			Pro MT		
	Avep	% chg	R-ret	Avep	% chg	R-ret
<i>Exespbase</i>	422	-	2163	393	-	2111
<i>Exespmono</i>	470*	+11.3%	2195	452*	+15.0%	2145
<i>Exesptcmono</i>	426*	+0.9%	2114	415	+5.6%	2081
<i>Exesptcqwgt</i>	372	-11.8%	1973	397	+1.0%	2039
<i>Exespcmqy</i>	462	+9.5%	2200	466	+18.6%	2148

Table 4 Retrieval results for topic translation for Spanish monolingual runs for both Sys and Pro MT, before and after applications of different feedback strategies.

Examination of Tables 1 to 4 reveals a number of consistent trends. Considering first the baseline runs. In all cases Sys MT translation of the topics produces better results than use of Pro MT. This is not too surprising since the documents were also translated with Sys MT, and the result indicates that consistency (and perhaps quality) of translation is important. All results show that our PRF results in improvement in performance over the baseline in cases. The variations in PRF results for query expansion for the different methods explored are very consistent. The best performance is observed in all cases, except Pro MT Spanish, using only the test collection for expansion term selection and collection weighting. Thus, although query expansion from pilot collections has been shown to be very effective in other retrieval tasks [6], the method did not work very well for CLEF 2003 documents and topics. Perhaps more surprising is the observation that term weight estimation from the pilot collection actually resulted in loss in average precision in most cases relative to the baseline. This result is very unexpected particularly since the method have been shown to be every effective and as been used with success in our past research work for CLEF 2001 and 2002.

Query expansion from the merged document collection (used for the multilingual task) of Spanish, English, French, and German also resulted in improvement in retrieval performance, in general slightly less than that achieved in the best results for French, German and Spanish using only the test collection. The result for this method is lower for Italian run, this is most certainly due to the absence of the Italian collection in the merged collection.

4.3 Bilingual runs

For the Bilingual task we submitted runs for Italian and Spanish tasks. Official runs are marked with a * and additional unofficial runs are presented. In all cases, results are presented for the following:

6. Baseline run without feedback (exebasebi)
7. Feedback runs using expanded query and term weights from the target collection (exebi)
8. Feedback runs using expanded query from pilot collection and term weights from test collection (exe*q+dtc)
9. Feedback runs using expanded query and term weights from pilot collection (exe*qd+tc)
10. We investigated further the effectiveness of pilot collection and the impact of vocabulary differences for different languages. This is done by expanding initial query statement from the topic collection and then applying the expanded query on the target collection (i.e. for German-Italian bilingual runs initial German query statement is expanded from the German collection and applied on the test collection) exe*q+dbi
11. Additionally both the expanded query and the corresponding term weight is estimated from the topic collection exe*qd+bi

Note: * and + refers to the either the topic or the target language e.g. sp -> Spanish, de-> German, it->Italian and fr->French. Results are presented for both Sys and Pro MT systems

4.3.1 Bilingual German to Italian

Run-id	Sys MT			Pro MT		
	Avep	% chg	R-ret	Avep	% chg	R-ret
<i>Exebasebi</i>	311	-	725	314	-	668
<i>Exebi</i>	370	+18.9%	748	359	+14.3%	701
<i>Exedeqitdte</i>	339	+9.0%	724	334	+6.4%	671
<i>Exedeqittdc</i>	327	+5.1%	715	335*	+6.7%	659
<i>Exedeqitd</i>	365	+17.4%	743	355*	+13.1%	691
<i>Exedeqitbi</i>	415*	+33.4%	750	397*	+26.4%	702

Table 5 Retrieval results for topic translation for Italian bilingual runs for both Sys and Pro MT, before and after applications of different feedback strategies.

4.3.2 Bilingual Italian to Spanish

Run-id	Systran MT			Globalink MT		
	Avep	% chg	R-ret	Avep	% chg	R-ret
<i>Exebasebi</i>	327	-	1938	349	-	1923
<i>Exebi</i>	376	+14.9%	2042	417	+19.5%	2064
<i>Exeitqspdtc</i>	331	+1.2%	1915	365	+4.6%	1940
<i>Exeitqdsptc</i>	339	+3.7%	1870	364*	+4.3%	1872
<i>Exeitqspd</i>	389	+18.9%	2071	417*	+19.5%	2011
<i>Exeitqspbi</i>	391*	+19.6%	2051	385*	10.3%	2004

Table 6 Retrieval results for topic translation for Spanish bilingual runs for both Sys and Pro MT, before and after applications of different feedback strategies.

For our bilingual run we tried a new method of query expansion and term weight estimation from the topic language collection. This resulted in the best performance for the Italian bilingual run with about 33% improvement in average precision. This method also worked well for the Spanish bilingual run giving about 19% improvement in average precision compared with results for baseline with no feedback. The standard method of query expansion and term weight estimation from the test collection also proved effective for the Italian-Spanish task. The use of term weights from the topic collection gives a large improvement over the result using test collection weights – positive in the case of the German-Italian task, but for the Italian-Spanish task this change has a negligible effect in the case of Systran MT and makes performance worse for Globalink MT. It is not immediately clear why these collections should behave differently, but it may relate to the size of the document collections, the Italian collection being much smaller than either of the German or Spanish collections.

Query expansion and term weight estimation from pilot collection resulted in improvement in average precision ranging from 1.2% to 9% for both results, although it failed to achieve comparable performance to other methods, which is again surprising but consistent with the monolingual results.

4.4 Multilingual Retrieval

Multilingual information retrieval presents a more challenging task in cross-lingual retrieval experiments, whereby a user submit a request in a single language (e.g. English) in order to retrieve relevant documents in different languages e.g. English, Spanish, Italian, German, etc. We approached this task in two ways. First, we retrieved relevant documents using the English queries individually from the four different collections and then merged the results together using different techniques (described below). Secondly we merged all the collections together to form a single collection and performed retrieval directly from this collection without using a separate merging stage.

Different techniques for merging separate result lists to form a single list have been proffered and tested. All of the techniques suggest that making assumptions that the distribution of relevant documents in the results set for retrieval from individual collection is similar is not true [15]. Hence, straight merging of relevant documents from the sources will result in poor combination.

Based on these assumptions we examined four merging techniques for combining the retrieved results from the four collections to form a single result list as follows:

$$u = \frac{doc_wgt}{g \max_wt * rank} \quad (3)$$

$$p = doc_wgt \quad (4)$$

$$s = \frac{doc_wgt}{g \max_wt} \quad (5)$$

$$d = \frac{doc_wgt - \min_wt}{\max_wt - \min_wt} \quad (6)$$

where u , p , s and d are the new document weight for all document in all collections and corresponding results are labelled $exemult4*$ where $*$ can be u , p , s or d depending on merging scheme used

doc_wgt = the initial document weight

$gmax_wt$ = the global maximum weight i.e the highest document from all collections for a given query

max_wt = the individual collection maximum weight for a given query

min_wt = the individual collection minimum weight for a given query

$rank$ = the a parameter to control the effect of size of collection, a collection with more document get a higher rank (value ranges between 1.5 and 1).

To test the effectiveness of the merging schemes, we merged all the four text collection into a single large combined collection. Expanded queries from this combined test collection ($exemultorg$) and from the TREC data pilot collection ($exemulttc$) were then applied on the resultant merged collection. For all official runs ($*$) English queries are expanded from the TREC-7 and 8 pilot collections and then applied on the test collection.

Run	Rec	P90	P70	%chg	Relevance
<i>Exemultbase</i>	351*	520	434	-8.4%	4574
<i>Exemult4u</i>	356*	532	438	-7.0%	4457
<i>Exemult4p</i>	356*	518	438	-7.0%	4428
<i>Exemult4s</i>	331*	525	433	-13.5%	4609
<i>Exemult4d</i>	438*	623	524	+14.3%	4828
<i>Exemulttc</i>	425	617	517	+10.9%	4853
<i>Exemultorg</i>	400	593	486	+4.4%	4675
<i>Exemult4snew</i>					

Note: an additional run $exemult4snew$ was conducted whereby the expanded query was estimated from the merged query collection and applied on the individual collection before merging using equation 5 above.

Table 8 Retrieval results for small Multilingual task before and after applications of different merging strategies.

The baseline result for our multilingual run ($exemultbase$) perhaps might not present a realistic platform for comparison with the feedback run using the different merging strategies ($exemult4*$). This is mainly because it was achieved from a no feedback run from the merged multilingual collection.

The multilingual results show that the different merging techniques provide similar retrieval performance. The result for merging strategy using equation 6 (which has been shown to be effective in past retrieval task) however resulted in about 14% loss in average precision compared to the baseline run. Also the merging strategies failed to show any improvement over raw score merging (row 3), although the merging strategy using equation 5, gave the highest number of relevant document retrieved for all the merging strategies.

Both our bilingual and monolingual runs show that retrieval results using expansion query and term weight estimation from pilot collection resulted in loss in average precision compared to baseline no feedback run in most cases. This might have contributed to the poor result from the different merging techniques for the multilingual runs (exemult4*). For the multilingual results using the merging techniques (exemult4*), We expanded the initial English query and estimated the term weights from the pilot collection and then applied these to the individual collections. However, results from our monolingual runs using this method were not very encouraging, and this might perhaps have contributed to the poor results after the application of the different merging techniques compared to the method whereby all the collections are merged to form one big collection.

To test this hypothesis, we conducted an additional run whereby we used the merged collection as the pilot collection and expanded the initial query from it. The expanded query was then applied on the individual collections and resultant result file merged using equation 5. The result showed an improvement of about 4% compared to that achieved from the baseline no feedback run from the merged collection (*Exemultbase*). It also resulted in about 11% increase in average precision over result from query expansion from the pilot collection (*Exemult4s*).

The best result for the multilingual task was achieved by expanding the initial query from the pilot collection and applying it on the merged collection. Query expansion from the merged collection (exemultorg) also resulted in about 10% improvement in average precision. These results suggest that merging a collection in a multilingual task might be more beneficial than merging the result lists taken from the retrieval from individual collections. This result is presumably due to the more robust and consistent parameter estimation in the combined document collection. In many practical situations combining collections in this way is not practical and multilingual IR can be viewed as distributed information retrieval task where there may be varying degrees of cooperation between the various collections.

5 Conclusions

For our participation in CLEF 2003 retrieval tasks we updated our synonym information to include common British and American English words. We explored the idea of query expansion from pilot collection and got some disappointing results which is contrary to past retrieval work utilizing the use of expanded queries and term weight estimation from pilot collections. This result may be caused by vocabulary and distribution mismatch between our translated test collection and the native English pilot collection, but further investigation is needed to ascertain whether this or other reasons underlie this negative result.

For the bilingual task we explored the idea of query expansion from a pilot collection in the topic language. This method resulted in better retrieval performance. Although we are working in English as our search language throughout this result is related to the ideas of pre-translation and post-translation feedback explored in earlier work on CLIR [2], and we need to perform further runs to explore possible further gains from the combination of both forms of feedback.

The different merging strategies used for combining our results for the multilingual task failed to perform better than raw score merging. Further investigation is needed to test these methods, particularly as some of them methods have been shown to be effective in past research. Merging the document collection resulted in better average precision than merging the result list. However, situations might arise whereby it is impossible to merge the various collections together, in this case an effective method of merging the result list is needed. Further investigation will be conducted to examine the possibility of improving the results achieved from merging result lists.

References

- [1] G.J.F. Jones, T. Sakai, N. H. Collier, A. Kumano and K. Sumita. A Comparison of Query Translation Methods for English-Japanese Cross-Language Information Retrieval. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 269-270, San Francisco, 1999. ACM.
- [2] L. Ballesteros and W. B. Croft. Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In Proceedings of the 20th Annual International ACM SIGIR conference on Research and Development in Information Retrieval, pages 84-91, Philadelphia, 1997. ACM.
- [3] G.Salton and C. Buckley. Improving Retrieval performance by Relevance Feedback. Journal of the American Society for Information Science, pages 288-297, 1990.
- [4] A.M. Lam-Adesina and G.J.F. Jones. Applying Summarization Techniques for Term Selection in Relevance Feedback. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1-9, New Orleans, 2001. ACM.

- [5] G.J.F. Jones and A.M. Lam-Adesina. Exeter at CLEF 2001: Experiments with Machine Translation for Bilingual Retrieval. In Proceedings of the CLEF 2001: Workshop on Cross-Language Information Retrieval and Evaluation, pages 59-77, Darmstadt, Germany, 2001.
- [6] S.E. Robertson, S. Walker, and M. M. Beaulieu. Okapi at TREC-7: automatic ad hoc, filtering, VLS and interactive track. In E. Voorhees and D.K. Harman, editors, Overview of the Seventh Text REtrieval Conference (TREC-7), pages 253-264. NIST, 1999.
- [7] M.F. Porter. An algorithm for suffix stripping. *Program*, 14:10-137, 1980.
- [8] S.E. Robertson, S. Walker, M. M. Beaulieu, M. Gatford, and A. Payne. Okapi at TREC-4. In D.K. Harman, editor, Overview of the Fourth Text Retrieval Conference (TREC-4), pages 73-96. NIST, 1996.
- [9] S.E. Robertson, S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 232-241, Dublin, 1994. ACM.
- [10] H.P. Luhn. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2):159-165, 1958.
- [11] H.P. Edmundson. New Methods in Automatic Abstracting. *Journal of the ACM*, 16(2):264-285, 1969
- [12] A. Tombros and M. Sanderson. The Advantages of Query-Biased Summaries in Information Retrieval. In proceedings of the 21st Annual International ACM SIGIR Conference Research and Development in Information Retrieval, pages 2-10, Melbourne, 1998. ACM.
- [13] S.E. Robertson. On term selection for query expansion. *Journal of Documentation*, 46:359-364, 1990.
- [14] S.E. Robertson, S. Walker. Okapi/Keenbow. In E. Voorhees and D.K. Harman, editors, Overview of the Eighth Text REtrieval Conference (TREC-8), pages 151-162. NIST, 2000
- [15] Jacques Savoy. Report on CLEF-2002 Experiments: Combining Multiple Sources of Evidence. In Proceedings of the CLEF 2002: Workshop on Cross-Language Information Retrieval and Evaluation, pages 31-46, Rome Italy, September 2002.