Query Expansion for Microblog Retrieval

Ayan Bandyopadhyay
ISI Kolkata

Mandar Mitra
ISI Kolkata

Prasenjit Majumder
DAIICT Gandhinagar

Abstract

Entries in microblogging sites such as Twitter are very short: a "tweet "can contain at most 140 characters. Given a user query, retrieving relevant tweets is particularly challenging since their extreme brevity exacerbates the well-known vocabulary mismatch problem. In this preliminary study, we explore standard query expansion approaches as a way to address this problem. Since the tweets are short, we use external corpora as a source for query expansion terms. Specifically, we used the Google Search API (GSA) to retrieve pages from the Web, and used the titles to expand queries. Initial results on the TREC 2011 Microblog test data are very promising. Since many of the TREC topics were oriented towards the news genre, we also tried restricting the GSA to a news site (BBC) in the hope that it would be a cleaner, less noisy source for expansion terms. This turned out to be counter-productive. Some analysis of these results is also included.

1 Introduction

Microblogging sites like `http://twitter.com` have emerged as a popular platform for expressing opinions. Given the increasing amount of information available through such microblogging sites, it would be nice to be able to retrieve useful tweets in response to a given information need. Finding relevant tweets that match a user query is challenging for the following reasons.

- Tweets are short. They contain a maximum of 140 characters.
- Tweets are not always written maintaining formal grammar and proper spelling. Spelling variations increase the likelihood of vocabulary mismatch.

In this preliminary study[1], we explore standard query expansion approaches as a way to address this problem.

Related work. Our approach is based on a method proposed by Kwok et al. [5] to improve weak ad-hoc queries through "web assistance", in which the Web (accessed via the Google search engine) is used as a source of expansion terms. We took the cue from this paper and used the Google Search API (GSA) to tap the Web as an external corpus for query expansion. Massoudi et al.[6] have proposed a language modeling approach to searching Microblog posts. Their method incorporates query expansion and uses certain "quality indicators" during matching. Hashtag retrieval [4] is also closely related to our work. Hashtags refer to certain important "keywords" in a message that are designated as tags using a hash (#) sign. Hashtags are useful as a very quick method for categorizing or tagging messages. Efron [4] showed that for a Twitter collection, hashtags can be predicted using query expansion. Dong et al. [3] have proposed a ranking method that takes both "relevance" and freshness into account. Del Corso et al. [2] also suggested a ranking method for news documents in which recency plays a major part.

[1] A version of this report has been submitted for possible publication at a refereed conference.

2 The data

We were able to download the HTML version of only 15,249,660 tweets (not the whole collection) within 9th September 2011 (run submission deadline was 11th September 2011). After the run submission deadline, some more tweets were downloaded. The final collection contains 16,087,002 tweets. The downloaded tweets were filtered using the following rules:

- removed tweets containing only punctuation;
- removed tweets for which 70% or more of the total content is part of a URL;
- removed tweets for which 20% or more of the total content is non-ASCII;
- removed those tweets for which the HTTP status is 403 or 404;
- removed re-tweets starting with "RT".
- tweets with HTTP status 302 but not marked "RT" were removed for the final corpus (FC), but not for submitted runs' (SR) corpus.

Table 1: Corpus Statistics

HTTP Status	**No. of tweets in the corpus used for submission**	**No. of tweets in final corpus**
200 (OK)	12,530,843	13,181,737
301 (Moved Permanently)	897,836	987,866
302 (Found)	1,004,562	1,054,459
403 (Forbidden)	377,657	404,549
404 (Not Found)	438,759	458,388
Unknown [2]	3	3
Total	**15,249,660**	**16,087,002**

Statistics related to the SR and FC are shown in table 1.

3 Our Approach

3.1 Basic Retrieval Approach

We have used the TERRIER IR system[3] for our experiments. Stopwords were removed and Porter's stemmer was used. The top 1,000 tweets were retrieved using the InL2c1 model [1] for each query. We then removed those tweets which were posted after the query tweet. Finally, we selected only the top fifty results per query for evaluation. The rationale for this step is the following.

According to the task definition, the retrieved list of tweets are ranked in order of time, newest to oldest, prior to evaluation. In other words, the final ranked list should maintain the fresh-tweet-first property. This (re)ordering creates an additional difficulty: determining the

[2]These 3 tweets are: "29516777562046464 302 null null", "33402257038909440 302 null null"and "33570228973604865 302 null null"

[3]http://terrier.org

number of tweets to return becomes crucially important. If a system retrieves a relevant stale tweet at the top, but a fresh non-relevant tweet at the end of the ranked list, the temporal re-ordering will cause the former to drop down the list and the latter to rise up. Thus, overall performance will suffer. This situation is more likely if the retrieved list is long. To handle this situation, we simply considered only the top 50 tweets per query (re-ranked according to recency) for evaluation.

As explained in the Introduction, the vocabulary mismatch problem is expected to be particularly severe for tweet retrieval. Our intention, therefore, was to explore query expansion (QE) approaches as a way to address this problem. Since the tweets are very short, standard techniques such as blind relevance feedback (where the documents themselves serve as a source of expansion terms) may not work well. Thus, we chose to use external corpora as a source for query expansion terms. Specifically, we used

- the Web, and
- the BBC news site (http://www.bbc.co.uk and http://www.bbc.co.uk/news/mobile).

3.2 Topic Processing for TREC 2011 Microblog submission

The original queries were submitted to the Google search API[4] (GSA). We took only the titles from the list of returned results. For each query, GSA returned a maximum of 8 pages of results, with a maximum of 8 results per page. Thus, at most $8 \times 8 = 64$ results were returned per query. The five most frequent word-level n-grams ($n = 1, 2, 3$) were added to the original topic. We also tried query *reformulation* (as opposed to QE) by excluding the original topics terms, and including only the n-grams obtained above. Details of the query processing steps for various submission to the TREC 2011 Microblog track are given below.

- **R1 (IRSIGoogle1G):** Results were retrieved for each query using the Google Search API. The title words ($n = 1$) from all the pages returned by Google were sorted in descending order by their frequencies. The most frequent five words were added to the original topic and retrieval was done using Terrier-3.5 with these new topics.
- **R2 (IRSIGoogle2G):** This is the same as R1, except that we use top five (word) 2-grams instead of single words.
- **R3 (Google1GNO):** This is the same as R1, except that we did not include the original topic terms during query expansion.
- **R4 (InL2c1):** Retrieval was done using the original queries provided by TREC.

Table 2: Example queries for submitted runs

Topic no	Original Query	Final Query				
		R1	R2	R3	R4	NBBCM1GQE[5]
14	release of "The Rite"	box dvd movie release rite release of "the rite"	720pbox date office release rite the release of "the rite"	box dvd movie release rite	release of "The Rite"	anglicans are british charles voice release of "the rite"

[4]http://code.google.com/apis/websearch

[5]Described in section 4.

3.3 Topic Processing for further experiments

We also tried restricting GSA to the BBC sites mentioned above. The top 10 valid / available documents from each site in turn were used for expansion. We repeated the previous process for expansion and reformulation. In addition to using the titles, we also tried using the content of the returned documents.

4 Result analysis

Tables 3 and 4 show how well runs R1, R2, R3, R4 did when compared to the best, median and worst figures. The columns marked SR (= submitted runs) correspond to our official submissions, while the columns marked FC correspond to the runs retrieved from the final collection (see section 2).
We use the following naming conventions for various retrieval runs:

- NBBCM1GQE — queries created through QE based on results from BBC's mobile site; queries were expanded using the five most frequent 1-grams from the titles of the returned results.
- MBA (resp. MMeA and MWA) denotes the mean of the best (resp. median and worst) average precision figures from all submissions (from all participants) to the TREC 2011 Microblog track.
- The Baseline run denotes the run provided by the TREC 2011 Microblog track coordinators.
- PRFB — result of running the default pseudo-relevance feedback method provided by TERRIER using the original topics, and the TREC 2011 Microblog corpus.

Table 3: Comparison of runs R1, R2, R3, R4

Run Name	**allrel**				**highrel**			
	P@30		**MAP**		**P@30**		**MAP**	
	SR	**FC**	**SR**	**FC**	**SR**	**FC**	**SR**	**FC**
R1	0.3347	0.4054	**0.2274**	0.2732	**0.1313**	0.1020	**0.2186**	0.1732
R2	0.3020	0.3673	0.2044	0.2583	0.1020	0.0857	0.1952	0.1550
R3	**0.3401**	**0.4143**	0.2265	**0.2740**	0.1232	**0.1041**	0.2114	**0.1739**
R4	0.2544	0.3156	0.1525	0.1989	0.0818	0.0741	0.1389	0.1224
Baseline	**0.0986**		**0.1411**		**0.0265**		**0.1089**	
MBA	**0.6116**		**0.5127**		**0.2646**		**0.3192**	
MMeA	**0.2592**		**0.1433**		**0.0687**		**0.0930**	
MWA	**0.0000**		**0.0000**		**0.0000**		**0.0000**	

Table 4: Comparison with respect to the baseline run (R4) by P@30, MAP of allrel measure

	Baseline	**R1**	**R2**	**R3**	**R4**
P@30	0.0986	0.4054 (+311.16%)	0.3673 (+272.52%)	0.4143 (+320.18%)	0.3156 (+220.08%)
MAP	0.1411	0.2732 (+93.62%)	0.2583 (+83.06%)	0.2740 (+94.19%)	0.1989 (+40.96%)

Table 5: Result for allrel (49 topics)

Run Name	Greater than or equal to best				Less than best but greater than median				Less than equal to median but greater than worst				Equal to worst			
	P@30		AP		P@30		AP		P@30		AP		P@30		AP	
	SR	FC	SR	FC	SR	FC	SR	FC	SR	FC	SR	FC	SR	FC	SR	FC
R1	4	4	1	1	30	35	34	41	14	10	13	7	1	0	2	0
R2	2	4	1	0	31	32	34	39	15	13	14	10	1	0	1	0
R3	3	4	1	1	32	34	35	40	12	10	13	7	2	1	1	1
R4	0	1	1	1	23	30	25	33	25	17	22	14	3	0	2	1

Table 6: Results for highrel (33 topics)

Run Name	Greater than or equal to best				Less than best but greater than median				Less than equal to median but greater than worst				Equal to worst			
	P@30		AP		P@30		AP		P@30		AP		P@30		AP	
	SR	FC	SR	FC	SR	FC	SR	FC	SR	FC	SR	FC	SR	FC	SR	FC
R1	4	24	2	5	19	7	22	25	6	2	8	3	4	0	3	0
R2	2	20	2	3	16	11	19	23	8	2	10	7	7	0	4	0
R3	4	24	2	5	17	6	22	24	7	2	9	3	5	1	2	1
R4	1	14	2	3	11	13	12	22	14	5	14	7	7	0	7	0

It is clear from Table 7 that query expansion and reformulation results in a huge improvement over the baseline run.

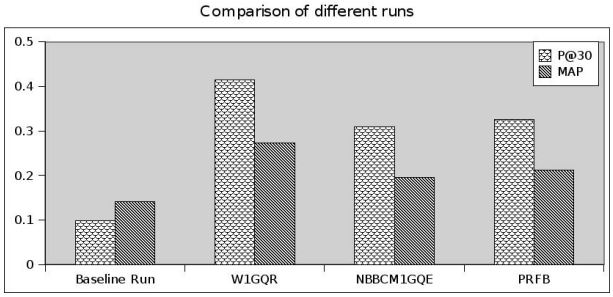
Table 7 compares the best runs obtained using different domains "the web", and the news domain ("http://www.bbc.co.uk"and "http://www.bbc.co.uk/news/mobile").

Table 7: Comparison of different runs

Run Name	P@30	MAP
Baseline run	0.0986	0.1411
PRFB	0.3252 (+229.82%)	0.2130 (+50.96%)
R3	**0.4143** (+320.18%)	**0.2740** (+94.19%)
NBBCM1GQE	0.3102 (+214.60%)	0.1964 (+39.19%)
MBA	0.6116	0.5127
MMeA	0.2592	0.1433
MWA	0.0000	0.0000

[6]The best value and the median value are equal for topic no 18, for both allrel AP and P@30 (Table 5). For allrel AP topic 15 has the same worst value and median value (Table 5), for allrel P@30, topics numbered 15 and 33 have equal worst and median values. For highrel P@30 measure (Table 6), topics numbered 16 and 18 have the same best and median. Whereas, topics numbered 14, 15, 23, 24, 26, 27, 28 and 32 have the same worst and median values. For highrel AP measure (Table 6), topics numbered 16 and 18 have the same best and median values, while, topics numbered 15 and 23 hold equal worst and median values.

Figure 1: Comparison with respect to the baseline run by P@30 and MAP



For the TREC 2011 Microblog track, the primary evaluation measure was P@30. The best P@30 figure over all submissions was reported to be 0.4551. We obtained a highest P@30 value of 0.4143 (run R3). This would be ranked 7th among all submissions. This run (R3) also achieved the best MAP score among all our runs (MAP = 0.2740). For comparison, the best MAP reported at TREC 2011 is 0.3350. On the basis of MAP, R3 would be ranked 6th.

It is clear from Table 2 that query expansion and reformulation using external resources results in significant improvements over the baseline run. Also, since many of the topics were from the news genre, we tried narrowing the scope of GSA to news sites (specifically BBC, as mentioned above). Unfortunately, the results were not at all impressive. NBBCM1GQE, the best among these runs, yielded a fairly good P@30 score, but the corresponding MAP was unsatisfactory (lower than even the PRFB run).

On the positive side, the improvements yielded by R3 and NBBCM1GQE runs over baseline run were found to be statistically significant (p-value = 3.827e-06 and p-value = 0.02077 respectively) by Wilcoxon signed-rank test.

5 Error Analysis

We looked at some queries for which NBBCM1GQE does significantly worse than R3. Table 2 shows one such query (topic no. 14). The original query asked for information related to the release of the film "The Rite". For R3, the new query after reformulation was "box dvd movie release rite". It is clear that the three new terms added to the query during the reformulation process are related to films. In contrast, for NBBCM1GQE, the expanded query was "anglicans are british charles voice release of the rite". The addition of a number of unrelated terms clearly destroys the focus of the query. As a result, the average precision (resp. P@30) for this topic drops from 0.0759 (resp. 0.5000) for R3 to 0.0006 (resp. 0.0333) for NBBCM1GQE.

Further, 1732 tweets that were judged by the assessors for the Microblog track were null tweets in the corpus crawled by us. Of these, 75 were judged relevant. If those relevant tweets were available in our corpus, it is possible that the performance of our runs would improve further.

6 Conclusion

Our approach for the TREC 2011 Microblog track was to expand / reformulate queries using external resources, viz. the Web, with the help of GSA. Our results were promising. We also narrowed down the search space to a news domain (BBC), as most of the queries were oriented towards the news genre. Unfortunately, QE and query reformulation using the news domain

produced worse results. Further analysis is needed to find out what went wrong. Also, further work is needed in order to explore different ways for expanding / reformulating queries.

References

[1] Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20:357–389, October 2002.

[2] Gianna M. Del Corso, Antonio Gulli, and Francesco Romani. Ranking a stream of news. In *in WWW 05: Proceedings of the 14th international conference on World Wide Web*, pages 97–106. ACM Press, 2005.

[3] Anlei Dong, Yi Chang, Zhaohui Zheng, Gilad Mishne, Jing Bai, Ruiqiang Zhang, Karolina Buchner, Ciya Liao, and Fernando Diaz. Towards recency ranking in web search. In *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, pages 11–20. ACM, 2010.

[4] Miles Efron. Hashtag retrieval in a microblogging environment. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 787–788, 2010.

[5] Kui-Lam Kwok, Laszlo Grunfeld, and Peter Deng. Improving weak ad-hoc retrieval by web assistance and data fusion. In Gary Geunbae Lee, Akio Yamada, Helen Meng, and Sung-Hyon Myaeng, editors, *AIRS*, volume 3689 of *Lecture Notes in Computer Science*, pages 17–30. Springer, 2005.

[6] K. Massoudi, E. Tsagkias, M. de Rijke, and W. Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. *ECIR 2011: 33rd European Conference on Information Retrieval*, pages 362–367, 2011.