# Overcoming Vocabulary Limitations in Twitter Microblogs

Yubin Kim
Language Technologies Institute
Carnegie Mellon University
yubink@cs.cmu.edu

Reyyan Yeniterzi
Language Technologies Institute
Carnegie Mellon University
reyyan@cs.cmu.edu

Jamie Callan
Language Technologies Institute
Carnegie Mellon University
callan@cs.cmu.edu

## ABSTRACT

One major difficulty in performing ad-hoc search on microblogs such as Twitter is the limited vocabulary of each document due their short length. In this paper, two approaches to addressing this issue are presented. The first is query expansion through pseudo-relevance feedback and the other is document expansion of tweets using web documents linked from the body of the tweet. Tweets are expanded by concatenating the contents of the title tag and the meta descriptor tags of the document to the tweet itself. These two approaches gave additive gains in MAP and Precision at 30.

## General Terms

Twitter, document expansion, query expansion, pseudo-relevance feedback

## 1. INTRODUCTION

In recent years, a new form of publishing has become increasingly popular on the Internet; in addition to longer articles, users are now publishing short messages called microblogs. One of the most popular microblog platforms is Twitter[1] and its users alone generate 400 million "tweets" per day as of June 2012[2]. These microblog entries have different characteristics from a typical web document; the entries are shorter due to a 140 character limit, are often filled with accidental and deliberate spelling errors, and they are real-time with new entries being constantly created. Therefore, techniques to address vocabulary mismatch and a timely real-time indexer become important for microblogs.

In this paper, the first need is considered within the framework of the Microblog Track in the Text REtrieval Conference[3] (TREC) and document expansion of tweets and query expansion through pseudo-relevance feedback are proposed as solutions. Phrases such as proper nouns extracted from queries are also investigated due to their significance in the query.

The real-time ad-hoc task in Microblog Track specifies that given a query which is issued at a specific time point, tweets that contain relevant information up to the specified time should be returned. Any retweets or tweets from a later time than which the query was issued are considered non-relevant.

To comply with Twitter's terms of use, the dataset for the Microblog Track was distributed as a list of tweet ids, which the participants used to fetch the actual tweet text from Twitter directly. However, because the full JSON form of tweets are not easily available without an API key, the following work was done on a corpus of tweets constructed by screen-scraping the tweet display HTML pages.

## 2. RELATED WORK

In the inaugural year of the Microblog Track at TREC, there were some common themes among participants. The first theme was query expansion. Tweets are short, thus there is often a vocabulary mismatch between a query and the content of the tweet. For example, the query "Detroit Auto Show" has a relevant tweet which contains "North American Car of the Year" but no direct mention of "auto". Although the exact implementation of their methods differed, all of the top 5 finishing runs included some form of query expansion [8, 1, 6, 9, 4]. Most reported that query expansion improved their results, although Louvan et al. saw that one of their query expansion methods hurt results for highly relevant tweets while a different method improved results for highly relevant tweets [7].

Another popular theme was temporal distance or recency, either as a feature in a learning-to-rank algorithm [8] or used as a boosting factor to re-rank tweets [1, 4, 10, 7]. However, the results of these experiments were mixed. Metzler et al. reported that the time feature received a 0 weight after training [8] and Ferguson et al. saw a negative impact on runs judged against all relevant documents [4]. However, Ferguson mentioned that a temporal re-weighting helped when the run was judged against only highly relevant documents [4] and Roegiest et al. [10] and Amati et al. [1] reported small gains when a temporal element was used in ranking.

[1] http://twitter.com
[2] http://www.huffingtonpost.com/2012/06/12/garys-social-media-count_n_1590113.html
[3] http://trec.nist.gov/

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| **NOV 2012** | | **00-00-2012 to 00-00-2012** |

| 4. TITLE AND SUBTITLE | 5a. CONTRACT NUMBER |
|---|---|
| **Overcoming Vocabulary Limitations in Twitter Microblogs** | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Carnegie Mellon University,Language Technologies Institute ,Pittsburgh,PA,15213** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT

**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

**Presented at the Twenty-First Text REtrieval Conference (TREC 2012) held in Gaithersburg, Maryland, November 6-9, 2012. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA). U.S. Government or Federal Rights License**

14. ABSTRACT

**One major di culty in performing ad-hoc search on mi- croblogs such as Twitter is the limited vocabulary of each document due their short length. In this paper, two ap- proaches to addressing this issue are presented. The rst is query expansion through pseudo-relevance feedback and the other is document expansion of tweets using web documents linked from the body of the tweet. Tweets are expanded by concatenating the contents of the title tag and the meta descriptor tags of the document to the tweet itself. These two approaches gave additive gains in MAP and Precision at 30.**

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT **unclassified** | b. ABSTRACT **unclassified** | c. THIS PAGE **unclassified** | **Same as Report (SAR)** | **8** | |

Some groups also used custom text-scoring functions with parameters that would be unusual for an ad-hoc web task. For example, Ferguson et al. used the Okapi BM25 model with parameters $k1 = b = 0$, effectively using binary term weighting and eliminating document length normalization [4]. Louvan et al. also modified the built-in Lucene scoring function to use binary term weighting in conjunction to temporal boosting.

# 3. INITIAL DATA PROCESSING

The guidelines in TREC task indicate that any tweets that are retweets (a tweet originally written by a different author that was forwarded) or are non-English tweets are non-relevant. Additionally, document expansion requires detecting and downloading the links included in the tweet. Thus, some pre-processing was necessary before the tweets could be indexed.

## 3.1 Tweet Body Cleaning

Tweets have a loose structure that is used to represent information other than the text body. For example, the @ symbol is used with a username when a tweet is directed towards another user (e.g. *@bob123 Your BBQ was awesome!*) and the # symbol is used to indicate keywords or topics in the tweet (e.g. *A big hurrah for physics with existence of #Higgs confirmed*). Tweets may also contain links to external pages with more details about the subject of the tweet.

Although these details could be used to improve search results (e.g. using the external links for document expansion), indexing them along with the text body hurts results. Due to the 140 character limitation on tweets, these structural details become a large percentage of the content when left in. Indeed, when the @ username mentions and URLs were stripped from the tweets, search results improved in preliminary tests with the queries from last year's track. However, hash tags provide almost a keyword-like summary of the tweet and thus contains valuable content. Therefore, each tweet was stripped of @ username mentions and URLs, but the hash tags phrases were left in.

## 3.2 URL Detection and Download

URLs were extracted from tweets using two simple regular expressions:

- `(\s|^)(www\.\w\S*[^[:punct:]\s])[[:punct:]]*`
- `(http://\w\S*[^[:punct:]\s])[[:punct:]]*`

For each URL found in the tweet corpus, the webpage was downloaded and saved to disk annotated with the tweet ID the URL was extracted from. These webpages were then used later for document expansion.

## 3.3 Retweet Detection

Ordinarily, retweets are marked by a field in the full JSON representation of tweets available with a Twiter API key. However, some retweets are "old-style" retweets where the author manually copies and pastes a tweet annotated with a RT. Also, in the case of the TREC dataset, only a limited, screen-scraped representation of tweets are available. After a few experiments, it was found that when the HTML page of a tweet returns a 302 status code it indicates that tweet is a (new-style) retweet. In order to detect old-style retweets, tweets that have the word "RT" near the beginning of the tweet (where "beginning" is heuristically determined to be the first 8 characters) were marked as retweets as well.

## 3.4 Language Identification

The Microblog Track guidelines stipulate that non-English tweets are non-relevant. Therefore, the tweets needed to be tagged with their language ID so that non-English tweets could be discarded.

The language of each tweet was identified using a C implementation of Textcat[4]. The standard pre-packaged language models were used for Chinese, Japanese, Korean, Portuguese, Arabic, and Russian and their various encodings. The language models for English, Spanish, French, German, and Dutch were replaced with models specifically trained for Twitter created by Carter et al. [2]

A tweet was first stripped of mentions of usernames and URL-like strings, then if it was too short it was replicated until it contained 25 characters. Note that any tweet with fewer than 8 characters was discarded for not having enough content; this was done because extremely short tweets in all likelihood cannot satisfy the informational needs of a query.

Afterwards, the cleaned tweet was passed to Textcat and any tweets that had English as one of the possible language ID tags (i.e. tweets that are maybe English) were included in the index. Experiments were also run using tweets that contained English as the *only* language tag (i.e. tweets that are certainly English), but using the looser, maybe-English approach yielded better results in the query set from last year. This is likely due to the fact that the stricter, certainly-English approach can miss actual English tweets, while non-English tweets allowed by the looser approach are likely filtered out by the query itself, which is in English.

# 4. PROPOSED APPROACH

Two approaches for overcoming vocabulary mismatch were attempted to create a competitive entry for the TREC Microblog Track.

The first method is the use of document expansion. The Microblog Track guidelines states that a tweet which links to a relevant webpage is considered relevant by the judges—that is, a linked webpage is considered an extension of the tweet. Therefore, document expansion was used to expand the tweet with relevant terms from a document it links to.

In addition to document expansion, a query expansion method was tried to overcome the vocabulary mismatch problem. Pseudo-relevance feedback (PRF), a well-known and effective method of query expansion was used to expand the query with terms from the retrieved relevant documents.

Finally, a heuristic procedure was used to extract phrases from query topics, which were given additional weight in

[4]http://software.wise-guys.nl/libtextcat/

| Original Query | Extracted Phrases |
|---|---|
| BBC World Service staff cut | BBC World Service |
| Oprah Winfrey half-sister | Oprah Winfrey<br>half-sister |
| release of "The Rite" | The Rite |
| Thorpe return in 2012 Olympics | Thorpe<br>2012<br>Olympics |
| Michelle Obama's obesity campaign | Michelle Obama |
| Kings' Speech awards | Kings' Speech |

Table 1: Example extracted phrases

```
#weight( 0.05 #combine(#3(bbc world service))
         0.95 #combine(bbc world service staff cuts)
       )
```

Figure 1: Phrase query

the final queries.

## 4.1 Phrase Detection

In baseline query experiments with the 2011 topics, it was noticed that a few queries did extremely poorly despite having many documents that were judged relevant in the pool. One such query was query 14 *release of "The Rite"*. Most tweets judged relevant for this query contained the phrase "The Rite" in contrast to the non-relevant tweets returned by the baseline system which often contained only the term "rite" without the "the".

One way to resolve similar issues is to include stopwords in the query and the indexed documents; however, this may not generalize well for phrases which do not contain stopwords yet still appear in quotes. Therefore, in addition to retaining stopwords, phrases were extracted from queries by using capitalization, quotation marks, or dashes. With the help of these heuristics, phrases such as proper names which are important keywords in queries could be extracted and searched as a phrase with additional weights. Furthermore, numbers in queries were also extracted and given additional weight due to their significance. For instance, in query 2 *2022 FIFA soccer* the *2022* year helps to filter out relevant tweets from any *FIFA* and *soccer* related tweets. Several example queries and the extracted phrases are given in Table 1.

Once the phrases were extracted, they were given additional weight and were used to construct a query similar to Figure 1.

## 4.2 Document Expansion

It was observed that relevant, informational tweets often contain a URL link. This is usually because it is difficult to convey a lot of information in a 140 characters (which is the limit for tweets), so users often tweet a headline of a news article and link to the body of the article. Thus, the tweets were expanded with the content of the documents that it linked to in order to overcome the vocabulary mismatch.

| Original Query | Expanded Terms |
|---|---|
| 2022 FIFA soccer | fifa, cup, 2022, world, qatar, held, stadium, winter, care, child |
| Egyptian curfew | curfew, egypt, defy, build, besiege, govern, mubarak, extend, street, demonstrate |
| Moscow airport bombing | moscow, bomb, airport, suicide, terrorist, reuter, busiest, kill, chechen, rebel |

Table 3: Example PRF expansion terms

Initially, document expansion was attempted by selecting the top-k TF-IDF weighted terms in a HTML document. However, this produced terms that were very poor in quality; selected words were very rare terms such as usernames and misspelled words that occur a few times in a single document but nowhere else. Attempts at heuristically correcting the problem such as discarding words that only occur in a single document did not improve the results. Similar problems existed with using Kullback-Leibler divergence between a background language model and the HTML document as the term weights.

Therefore, instead of algorithmically trying to determine which terms are most representative of the webpage, the summarization efforts made by the creators of the HTML pages were leveraged by extracting the contents of the `title` tag, and the contents of the `keywords` and `description` type `meta` tags. The text from these sources were added to the tweet as expansion terms. Although simple, this method has the advantage of using keywords that humans have identified as being a good summary of the document.

A comparison of the three document expansion methods is presented in Table 2.

## 4.3 Pseduo-Relevance Feedback

Pseudo-relevance feedback (PRF) is a typical query expansion method which assumes the top N retrieved documents are relevant, identifies discriminative terms from these documents and adds them to the original query as expanded terms. Indri has a built-in PRF method which is based on Lavrenko's Relevance Model [5].

In this model, a language model is built from the top retrieved documents, and terms from that language model are ranked by their weights. A weighted unigram query is then built with those top terms, and this query is combined with the original query to retrieve the final set of documents.

Different sets of pseudo-relevance feedback parameters were tried and the best results were obtained with the top 5 retrieved documents, adding top 10 feedback terms to the query and using a weight of 0.2 on the original query and 0.8 on the expanded query. Example queries and their expanded terms are given in Table 3.

# 5. EXPERIMENTAL SETUP

| Original Tweet | TF-IDF | KL Divergence | Metadata |
|---|---|---|---|
| Stanley Ho Gives Up SJM Stake | shai parcele oster sjm okeeffe 1257 tussle too568 rind | parcele 1257 shai oster too568 rind santorum648 tussle | Plans to divide nearly all of Macau gambling .... |
| I make about $30 a day via twitter | 1440 credite clickbank grav tid generator xml referr refresh | 1440 credite tid grav generator xml clickbank referr refresh | make money online, affiliate marketing, make money on internet .... |
| jetsbuzztap: Yahoo! Sports > > Jets silenced: Steelers win AFC title 24-19 : jetsbuzzta.. | buzztap jetsbuzztap i5lx4 silenc retweete afc -19 silence steeler | jetsbuzztap i5lx4 buzztap silenc retweete -19 afc silence footer | By BARRY WILNER AP Pro Football Writer - National Football League news |

Table 2: Sample expansion terms for tweets

| Tweet Type | Number of Tweets |
|---|---|
| HTTP 200 | 11,796,107 |
| HTTP 301 | 2,212,522 |
| Non Retweets | 11,286,497 |
| English | 6,578,488 |
| Tweets Indexed | 5,776,034 |

Table 4: Statistics of tweets used

## 5.1 Statistics

The tweet corpus was downloaded on October 13, 2011. However, the last ID file was incompletely processed and was re-crawled on March 20, 2012. The statistics for the downloaded tweets are presented in Table 4.

HTML documents linked from non-retweet, English tweets were crawled over a few days with the majority being downloaded between March 14, 2011 to March 16, 2011. However, the webpages linked from the tweets from the last ID file which was not completely crawled initially, were downloaded on April 2, 2011. In total, 1,161,041 HTML documents were downloaded.

Of the tweets downloaded, many were found to be either retweets or non-English tweets. After these tweets were filtered out, only 5,776,034 tweets remained and were indexed.

## 5.2 Document Format

The indices for the dataset were created using Indri, a search engine commonly used in the research community[5].

With the initial data processing steps completed, each tweet was converted to a pseudo-XML "trectext" format suitable for indexing with Indri. Document expansion was done at this step; for each tweet to be converted to XML, the tweet ID was matched with the webpages previously downloaded as described in Section 3.2 and the related terms identified by the document expansion algorithm were added to the tweet content. An example of a trectext document can be seen in Figure 2.

The `CLEANTWEET` field is the original tweet cleaned as per Section 3.1. `EXPAND` contains the document expansion terms and

[5]http://www.lemurproject.org/indri/

```
<DOC>
<DOCNO>28965141634613248</DOCNO>
<TWEETID>28965141634613248</TWEETID>
<CLEANTWEET>Check our FB promo to #win #free stuff:
Grand Prize = #ski package to any resort in
Nth America? #ski #snowboarding</CLEANTWEET>
<EXPAND>Promotions on Facebook | Facebook  </EXPAND>
<EXPANDTWEET>Promotions on Facebook | Facebook
Check our FB promo to #win #free stuff:
Grand Prize = #ski package to any resort in
Nth America? #ski #snowboarding</EXPANDTWEET>
<DATE>Sun Jan 23 00:00:02 +0000 2011</DATE>
<USER>stormsale</USER>
<ONLYENGLISH>true</ONLYENGLISH>
</DOC>
```

Figure 2: Trectext format document

`EXPANDTWEET` is a concatenation of `EXPAND` and `CLEANTWEET`. `ONLYENGLISH` is true when a tweet is certainly-English, false if it is maybe-English (non-English tweets are not indexed). For more details regarding language identification, refer to Section 3.4.

In order to comply with the track requirement that no future data is used (even for collection statistics), an index was created for each topic which only contains the tweets that are in the "past" relative to the query, so that the IDF values are unpolluted by "future" documents upon query time.

## 5.3 Individual Component Results

To discern the individual effects of the different system components, the queries and relevance judgements from the 2011 Microblog Track were used to evaluate each component.

The official metric for the 2011 Microblog Track was Precision at 30. However, this year the track switched to the area under the receiver operating characteristic (ROC) curve. The ROC curve is a plot of the false positive rate (non-relevant retrieved) the true positive rate (relevant retrieved), and the area under the curve (AUC) is the probability that a classifier will rank a randomly chosen positive instance higher than a random negative instance [3]. In this paper, both metrics and Mean Average Precision (MAP) are reported.

A summary of the individual effects of the different components is shown in Table 5. Also, a per-query Average Precision (AP) analysis for four of the runs is shown in Figure 3. In all experiments, unless otherwise indicated, stopwords are included in all queries and the Dirichlet smoothing parameter is set to $\mu = 400$ as these settings produced the best results on training data. Each ranked list returns 1000 results.

As expected, pseudo-relevance feedback increases the average performance but with a higher variance in performance (compare standard deviations 0.2195 for baseline and 0.2568 for PRF). Document expansion also increased average performance to a lesser degree than PRF, but it *decreased* the standard deviation of the queries (0.2087). However, this effect may be incidental to the training query set, as the submitted runs did not exhibit similar effects.

When combined together, PRF and document expansion show statistically significant additive gains compared to running either PRF or document expansion alone.

Phrases did not produce any statistically significant results and the average gains are small. However, this is not surprising as stopwords are included these runs. In preliminary results, when phrase queries were compared with a stopped baseline, the gains for phrases were greatest in queries where important stopwords would have been removed from queries (such as *release of "The Rite"* or *release of "The Known and Unknown"*). The gains for these queries are less when compared to a baseline that includes stopwords.

Lastly, pre-processing of tweets is shown to significantly improve results; there is over a 10 point gain in both MAP and P@30 from an index of raw tweets to the `CLEANTWEET` index, which has non-English and retweets removed and the tweet body cleaned of URLs and user @ mentions.

# 6. EXPERIMENTAL RESULTS

Although the official metric this year was ROC, there was no single summary number given for the metric. Therefore, the results of the 2012 track are presented with MAP and Precision at 30.

Again, similar to the individual component experiments, in all of the runs, unless otherwise indicated, stopwords are included in all queries and the Dirichlet smoothing parameter is set to $\mu = 400$. Each ranked list returned 1000 results.

The following four configurations were chosen for the submitted runs.

1. cmuPhrE: Phrase query on document expansion field.
2. cmuPrfPhr: Phrase query linearly interpolated with PRF (on plain, cleaned tweets).
3. cmuPrfPhrE: Phrase query on document expansion field linearly interpolated with PRF performed on document expansion field.
4. cmuPrfPhrENo: Phrase query on document expansion field linearly interpolated with PRF performed on document expansion field where stopword PRF terms were removed.

The performance of the four runs in TREC 2012 are displayed in Table 6. For comparison, a baseline run index containing all the tweets in their raw form is also shown in the table. A per-query AP analysis is presented in Figure 4. The best run submitted (ordered by MAP and P@30), cmuPrfPhrENo, contained 45 out 59 queries above the median. Another run, cmuPrfPhr, recorded 46 queries above the median.

PRF continued to perform well in the 2012 query set, bringing statistically significant gains in both P@30 and MAP.

Document expansion did not work as well in the 2012 results. The differences in in cmuPrfPhr and cmuPrfPhrE seem to indicate a small additive average gain when document expansion and PRF are combined, at the expense of a higher variance. However, this gain is not statistically significant. This explains the '# above Median' column of Table 6, which shows that more queries were above median in cmuPrfPhr than cmuPrfPhrE. A possible cause for this may be the following.

The 2011 and 2012 query sets both had similar proportions of expanded relevant tweets over the total set of all relevant tweets, but the 2012 query set had a higher variance in the ratio of expanded tweets across different queries. That is, the 2011 query set had a more even distribution of relevant expanded tweets over different queries. This would make the effect of document expansion more variable over different queries in the 2012 query set, resulting in less statistically significant improvements.

The higher variance of the document expansion run compared to a run without expansion (cmuPrfPhr vs. cmuPrfPhrE) also differs from the findings from the 2011 query set, where document expansion was seen to reduce query performance variance from the baseline and when combined with PRF.

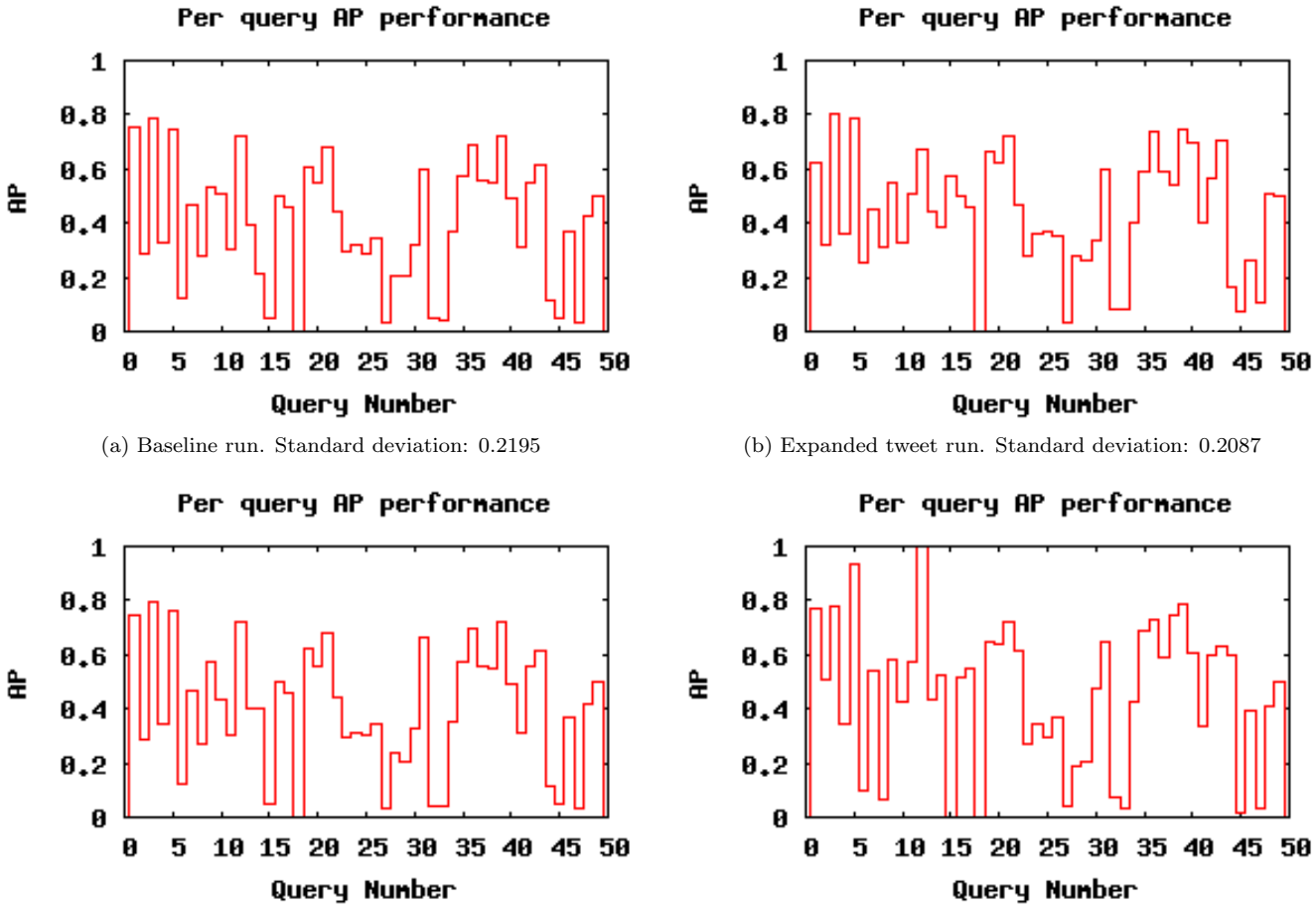
Unlike in 2011, the run without stopwords (cmuPrfPhrENo) did slightly better on average than the equivalent run including stopwords (cmuPrfPhrE) in the 2012 query set. Previously in the 2011 query data, *including* stopwords had slightly improved results. This is likely due to the 2011 query set containing more queries that would be sensitive to stopword removal, such as MB014 'release of "The Rite"' and MB018 'William and Kate fax save-the-date'.

# 7. CONCLUSIONS

To develop ad-hoc search techniques better suited for the new publishing medium of microblogs, two approaches were tried. The first approach performed document expansion using the webpages linked from the tweet in a pre-processing step. Two different methods of generating candidate expansion terms, TF-IDF and KL divergence, produced very similar and low-quality words. Using the title text and metadata from the webpages as the expansion terms offers an easy performance boost without delving into more complex natural language processing algorithms.

| | All relevant | | | Highly relevant | | |
|---|---|---|---|---|---|---|
| Method | AUC | P@30 | MAP | AUC | P@30 | MAP |
| Raw Tweet | 0.8535 | 0.3524 | 0.2894 | 0.8859 | 0.0769 | 0.1372 |
| `CLEANTWEET` | 0.8606 | 0.4619† | 0.3945† | 0.8976 | 0.1102† | 0.1894† |
| Phrase | 0.8614 | 0.4639 | 0.4011 | 0.8981 | 0.1102 | 0.1888 |
| Doc Exp | 0.8767* | 0.4837 | 0.4373* | 0.9090 | 0.1265* | 0.2200 |
| PRF | 0.8671 | 0.5088* | 0.4547* | 0.901 | 0.1197 | 0.2120 |
| Doc Exp + PRF | 0.8953* | 0.5116*ˆ | 0.4906*ˆ | 0.9201* | 0.1340* | 0.2362* |

Table 5: Effects of each individual component on TREC 2011 dataset.
† indicates a significant ($p < 0.05$) difference using the two-tailed paired t-test compared to the raw tweet.
* indicates a significant difference compared to `CLEANTWEET`.
ˆ indicates a significant difference compared to PRF and Doc Exp.




(a) Baseline run. Standard deviation: 0.2195

(b) Expanded tweet run. Standard deviation: 0.2087

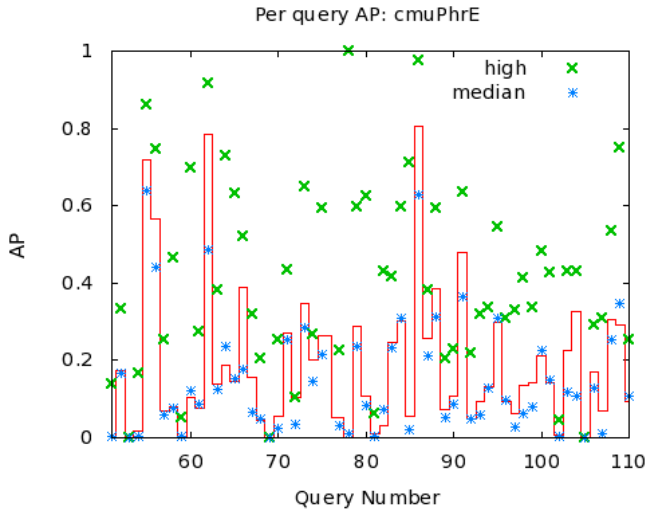
(c) Phrase run. Standard deviation: 0.2206

(d) PRF run. Standard deviation: 0.2568

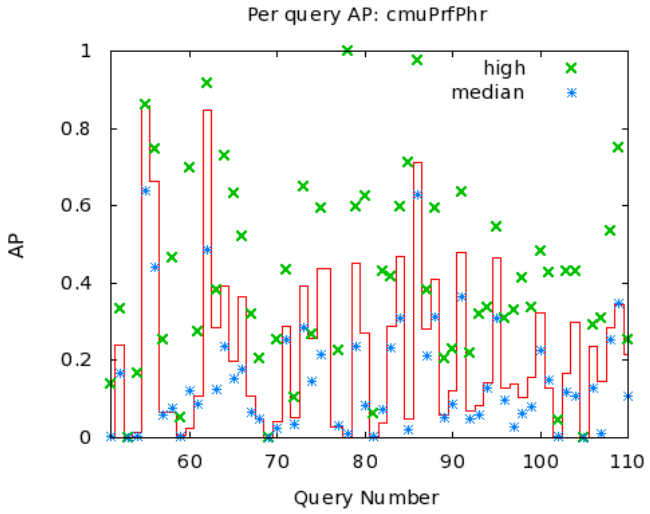
Figure 3: Per query analysis of AP values for all relevant tweets from TREC 2011 dataset.

| | All relevant | | |
|---|---|---|---|
| Run ID | P@30 | MAP | # above Median (out of 59) |
| Raw Tweet | 0.1605 | 0.1409 | 24 |
| cmuPhrE | 0.1966* | 0.1854* | 40 |
| cmuPrfPhr | 0.2266* | 0.2178* | 46 |
| cmuPrfPhrE | 0.2305 | 0.2200 | 44 |
| cmuPrfPhrENo | 0.2333 | 0.2223 | 45 |

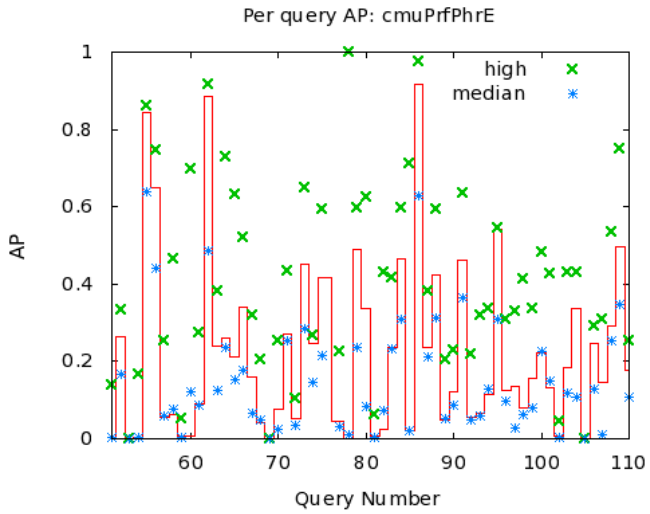
Table 6: Performance of four submitted official runs on TREC 2012 dataset. * indicates statistically significant ($p < 0.05$) difference between it and all previous configurations.



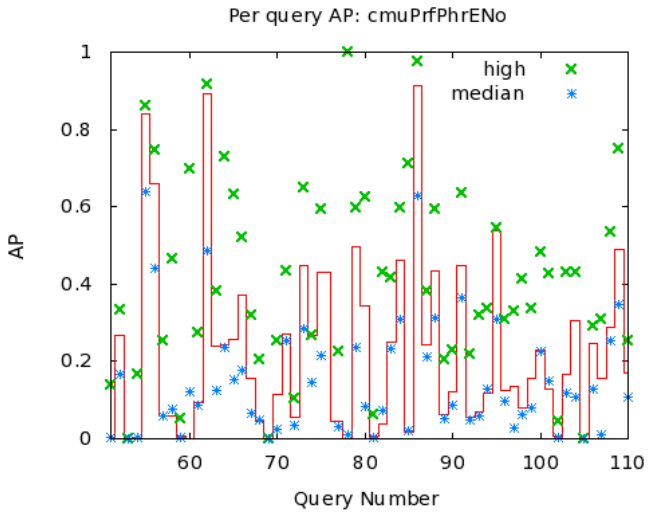

(a) cmuPhrE run. Standard deviation: 0.1853.




(b) cmuPrfPhr run. Standard deviation: 0.2083.




(c) cmuPrfPhrE run. Standard deviation: 0.2237.




(d) cmuPrfPhrENo run. Standard deviation: 0.2235.

Figure 4: Per query analysis of AP values for all relevant tweets on TREC 2012 dataset.

The second approach, in a run-time step, used pseudo-relevance feedback to perform query expansion which significantly improved average results at the cost of a higher variance.

When used together document expansion and query expansion gave additive gains in both the 2011 and 2012 query sets, although the gain was not statistically significant in the 2012 query set due to a higher variability of expanded documents marked relevant in different queries. Pre-processing the corpus to remove retweets and non-English tweets and cleaning the tweet body also gave a significant boost in results.

## 8. ACKNOWLEDGEMENTS

This research was in part supported by the National Science Foundation (NSF) grant IIS-0916553 and the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office. Furthermore this publication was made possible by the generous support of the iLab and the Center for the Future of Work. Any opinions, findings, conclusions, and recommendations expressed in this paper are the authors' and do not necessarily reflect those of the sponsors.

## 9. REFERENCES

[1] G. Amati, G. Amodeo, M. Bianchi, A. Celi, C. D. Nicola, M. Flammini, C. Gaibisso, G. Gambosi, and G. Marcone. FUB, IASI-CNR, UNIVAQ at TREC 2011. In *Text REtrieval Conference Proceedings*. NIST, 2011.

[2] S. Carter, W. Weerkamp, and E. Tsagkias. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation Journal*, 2012.

[3] T. Fawcett. Roc graphs: Notes and practical considerations for data mining researchers. Technical report, HP Laboratories Palo Alto, 2003.

[4] P. Ferguson, N. O'Hare, J. Lanagan, and A. F. Smeaton. CLARITY at the TREC 2011 Microblog Track. In *Text REtrieval Conference Proceedings*. NIST, 2011.

[5] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 120–127, New York, NY, USA, 2001. ACM.

[6] Y. Li, Z. Zhang, W. Lv, Q. Xie, Y. Lin, R. Xu, W. Xu, G. Chen, and J. Guo. PRIS at TREC2011 Microblog Track. In *Text REtrieval Conference Proceedings*. NIST, 2011.

[7] S. Louvan, M. Ibrahim, M. Adriani, C. Vania, B. Distiawan, and M. Z. Wanagiri. University of Indonesia at TREC 2011 Microblog Track. In *Text REtrieval Conference Proceedings*. NIST, 2011.

[8] D. Metzler and C. Cai. USC/ISI at TREC 2011: Microblog Track. In *Text REtrieval Conference Proceedings*. NIST, 2011.

[9] T. Miyanishi, N. Okamura, X. Liu, K. Seki, and K. Uehara. TREC 2011 Microblog Track Experiments at Kobe University. In *Text REtrieval Conference Proceedings*. NIST, 2011.

[10] A. Roegiest and G. V. Cormack. University of Waterloo at TREC 2011: Microblog Track. In *Text REtrieval Conference Proceedings*. NIST, 2011.