CLARITY at the TREC 2011 Microblog Track

Paul Ferguson, Neil O'Hare, James Lanagan, Alan F. Smeaton
CLARITY: Centre for Sensor Web Technologies,
Dublin City University.

Owen Phelan, Kevin McCarthy, Barry Smyth
CLARITY: Centre for Sensor Web Technologies,
University College Dublin.

Abstract

For the first year of the TREC Microblog Track the CLARITY group concentrated on a number of areas, investigating the underlying term weighting scheme for ranking tweets, incorporating query expansion to introduce new terms into the query, as well as introducing an element of temporal re-weighting based on the temporal distribution of assumed relevant microblogs.

1 Introduction

The introduction of the Microblog Track at TREC provides an opportunity to develop new research on a large corpus of microblog data (Tweets2011 collection). The CLARITY group took this opportunity to investigate a number of approaches to the retrieval of microblog data, which exhibits different characteristics than text traditional corpora.

In Section 2 we discuss the our baseline approach for ranking microblogs in response to a user query. Section 3 outlines our method for providing query expansion, while Section 4 describes our approach for re-weighting results based on their temporal proximity to a set of assumed relevant microblogs. Our experiments are described in Section 5 and we draw conclusions in Section 6.

2 Tweet Sorting

To provide a term-based document sorting baseline we used the Okapi BM25 model [1], which has been used extensively within the TREC community on a variety of corpora. The base formula that we used is as follows:

$$Score_{bm25}(q,d) = \sum_{t \epsilon q} log\left(\frac{N - df_t + 0.5}{df_t + 0.5}\right) \times \frac{(k_1+1)tf_t}{k_1((1-b)+b\frac{dl}{avdl}) + tf_t} \quad (1)$$

Here tf_t represents the within document term frequency, dl is the document length and $avdl$ is the average document length over the collection. Also, df_t is the number of times the term t occurs within the collection although, in order to comply with the TREC Microblog requirements for not using future data, this value was dynamically calculated based on the query time of the topic: effectively finding the number of occurrences of the term t from the start time of the collection until the time the query was issued.

Due to the short nature of microblogs we choose to set the BM25 parameters to effectively use a binary term weighting scheme within the BM25 model and, to discount document normalisation, to offer no penalty against longer microblogs. This can be achieved with

the BM25 model by setting the parameters $k1$ and b to 0. This simplifies Equation 1 so that only the *inverse document frequency* (IDF) component of the equation is considered, leaving us with the following:

$$Score_{bm25-idf}(q,d) = \sum_{t \epsilon q,d} log\left(\frac{N - df_t + 0.5}{df_t + 0.5}\right) \quad (2)$$

One of the drawbacks of simplifying the BM25 to that shown in Equation 2 is that there is a likelihood that a large number of results will have tied scores. We choose to resolve these ties was based on the recency of the microblogs, so that for any tied scores the most recent microblogs would appear first in the ranked list.

3 Query Expansion

Since microblogs are very short documents and the query topics for this task are short, the query terms may not always be present in a relevant document. For this reason, we use standard pseudo relevance feedback query expansion techniques to add new terms to the query, in an attempt to create a better representation of the topic.

Since our baseline approach ranks solely be the IDF component of the BM25 formula described in Equation 2, we can calculate the maximum possible ranking score for any document as the sum of the IDF scores for all of the query terms: any document containing all the query terms will have this maximum score. Rather than using the standard approach to pseudo relevance feedback and assuming that the top N microblogs are relevan, whether the value of N is chosen in advance, instead we assume relevance on the ratio of a microblog's score to this maximum possible score:

$$Score_{ratio} = \frac{Score_{bm25-idf}(q,d)}{max(Score_{bm25-idf}(q))} \quad (3)$$

Note that the maximum score in this formula refers to the maximum possible score, and not the score if the highest ranked document. If $Score_{ratio}$ is greater than a threshold value, λ, then a document is assumed to be relevant. In the absence of any ground truth *query relevance* (qrel) data to tune this parameter, the value chosen was, by necessity, somewhat arbitrary. For the experiments described here, a value of $\lambda = 0.7$ was used, meaning a document's relevance score needed to be more than 70% of the maximum possible score to be assumed relevant. This approach will cause N, the number of assumed relevant documents, to vary from topic to topic, and some topics will have no assumed relevant documents.

Once we have chosen a set of assumed relevant documents, we use the Robertson Selection Value [2] to rank the terms in this pseudo relevant set:

$$rsv = \frac{r}{N} \times rw \quad (4)$$

where r is the number of relevant document the term occurs in, rw is the IDF weighting of the term, and N is the number of assumed relevant microblogs. After ranking candidate terms by their rsv score, we add the top X terms to the query, and then use Equation 2 to rank microblogs by this new query. Again, we arbitrarily choose a value for X, setting it to 4.

4 Temporal Re-weighting

In order to test the hypothesis that the time a microblog was published is an important factor in determining its relevance to a given query, we use pseudo relevance feedback to model the temporal distribution of a topic. We use the approach described in the previous

section to create pseudo relevance judgements for a topic. If there are 2 or more assumed relevant microblogs for a topic, we make the initial assumption that the oldest of these represents the start time for a topic, and that the most recent of these represents the end time for the topic. We then expand this temporal extent for the topic as follows:

$$range_{init} = end\_time_{init} - start\_time_{init} \tag{5}$$

$$start\_time_{extented} = start\_time_{init} - \frac{range}{\alpha} \tag{6}$$

$$end\_time_{extended} = end\_time_{init} + \frac{range}{\alpha} \tag{7}$$

$$range_{extended} = end\_time_{extended} - start\_time_{extended} \tag{8}$$

Equation 5 calculates the temporal range of the topic as the amount of time between the first and last relevant microblog. We then expand this temporal range, with the extent of the expansion determined by the α parameter: a value of $\alpha = \infty$ would result in no expansion of a topic's temporal range, while a value of 2 would double the the temporal range of a topic (i.e. a 50% expansion on either side).

If a microblog falls outside of this temporal range, we re-weight microblogs based on this 'temporal centre' of the topic:

$$Score_{temporal} = Score_{bm25-idf}(q, d) \times temporal\_weight \tag{9}$$

where the temporal weight is calculated as follows:

$$decay\_factor = \frac{range_{extended}}{\alpha} \tag{10}$$

$$temporal\_weight = \frac{decay\_factor}{distance_{temp}} \tag{11}$$

The $distance_{temp}$ is the temporal distance of a microblog from the start/end of the topic's temporal range. The $decay\_factor$ controls how quickly a microblog's score decays as it's distance from the temporal centre of a query increases, and is controlled by the same α parameter used in Equations 6 and 7, with a larger parameter value enforcing a more severe temporal decay.

5 Experiments

For our four official submissions to the TREC Microblog Track we submitted the following runs: CLARITY1, CLARITY2, CLARITY3, CLARITY4, which we will now describe in more detail.

CLARITY1: this provides us with a baseline run, and does not use future data or external evidence. The run uses the BM25 sorting approach described in Equation 2 in Section 2 and, as described previously, tied scores are resolved based on the recency of the tweet, so that the most recent tweet appears higher in the ranked list.

CLARITY2: in addition to the baseline sorting provided by CLARITY1, the CLARITY2 run adds query expansion with pseudo relevance feedback as described in Section 3. The number of assumed relevant documents will vary from topic to topic, with the *relevance threshold*(λ) set to 70%. The number of query expansion terms is fixed at 4. This run does not use future data or external evidence.

CLARITY3: the CLARITY3 run uses the same approach as the previous CLARITY2 run, however in addition it uses a language classification step in an attempt to filter non-English tweets (which are considered to be non-relevant). For this language classifier we

used a Language Detection Library[1] which we use to filter out non-English tweets, prior to query expansion. Once again this run does not include any future data. Due due to the inclusion of the language detection component, however, we consider this to use external evidence.

CLARITY4: this builds upon the CLARITY3 run, but also adds a temporal re-weighting element (as described in Section 4) which downweights tweets that are far from the temporal centre of the assumed set of relevant tweets. For this run we set the temporal decay factor, α, to 2. Similar to CLARITY3 this run uses external data (due to the inclusion of the language classifier). As with all of our runs, no future data is used.

It is worth noting that for all runs, having produced a final ranked list (as described for each run above), finally we truncated each list to the top 30 most relevant tweets before these were re-ranked in reverse chronological order.

5.1 Analysis

For this year's Microblog track precision at 30 (P30) was chosen to be the main evaluation measure, and so for each run we report performance based on P30. Figures 1 shows the results for *all relevant* and 2 shows the results for *highly relevant* microblogs. For comparison purposes, we include a number of other baselines provided by TREC. The "best" run is a pseudo run that we created by taking the union of the top performing of all submitted runs for each topic. The "median" is also a pseudo run, this time comprising the median runs for each topic. Finally the "disjunctive baseline" consists of a single run which was generated using Lucene[2], selecting the most recent 1000 tweets that contain any of the query terms.

We can see that the baseline run CLARITY1 outperforms all other runs. For the both *all relevant* and *highly relevant* conditions, using query expansion harms performance. The CLARITY3 run, however, is an improvement over CLARITY2, showing a minor benefit from using language filtering. Comparing CLARITY4 with CLARITY3, we can see that temporal re-weighting harms performance for the *all relevant* condition. For the *highly relevant* condition, however, temporal re-weighting gives a small improvement, suggesting that temporal re-weighting approach may be beneficial, and is worth further investigation.

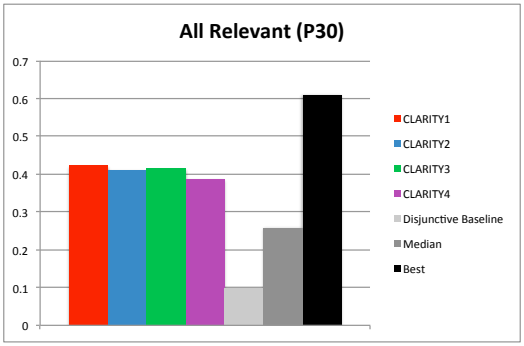

Figure 1: P30 scores for our official runs compared with best and median results from all participants as well as a TREC supplied disjunctive baseline (on all relevant results).

Figures 3 and 4 show the P30 performance on a topic by topic basis for *all relevant* and *highly relevant* tweets respectively. We can see that overall CLARITY1 performs best, although on certain queries the other runs gain higher scores, in particular for the *highly relevant* condition. For a number of queries the runs CLARITY2, CLARITY3 and CLARITY4 perform similarly or identically to the baseline. This is due of our pseudo relevance

[1] http://code.google.com/p/language-detection
[2] http://lucene.apache.org/java/docs/index.html

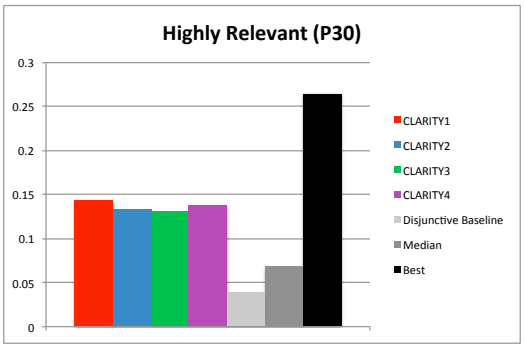

Figure 2: P30 scores for our official runs compared with best and median results from all participants as well as a TREC supplied disjunctive baseline (on highly relevant results).

feedback method, which will result in a different number of documents being assumed relevant for each topic. If no documents are assumed relevant for a given topic, then query expansion and temporal re-weighting will not be used, meaning there will be no difference from the baseline. This approach could be considered reasonably conservative and limits the influence of query expansion and temporal re-weighting, although without further investigation it is not clear whether an increased influence for these would increase or decrease the overall performance.

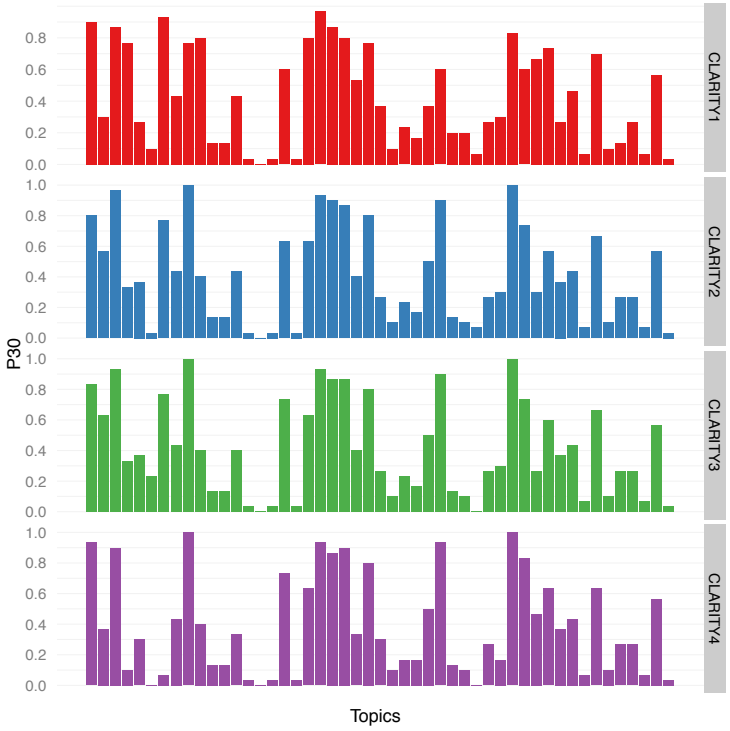

Figure 3: P30 performance for each run, on a topic by topic basis for all relevant results.

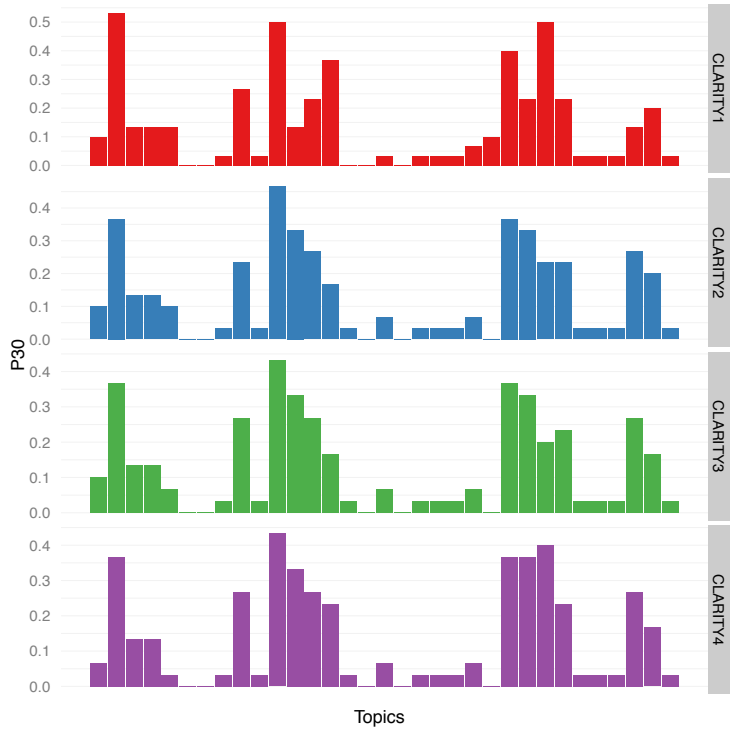

Figure 4: P30 performance for each run, on a topic by topic basis for highly relevant results.

6 Conclusions

In our experiments we investigated a number of different approaches for microblog retrieval. None of our runs used future data, while 2 of our runs use external data in the form of an external language detection library. Our baseline run CLARITY1 focused on the underlying term weightings by ignoring term frequency and document length information, in an attempt to produce a retrieval approach suitable to microblog data. This seems to have been quite successful, achieving high P30 scores for both the *all relevant* and *highly relevant* tweets. The incorporation of query expansion proved to be useful for certain queries, although overall it performed worse than the baseline run. Finally, the use of temporal re-weighting improves performance slightly for the highly relevant condition (CLARITY4 vs CLARITY3). Since temporal re-weighting was incorporated alongside with our query expansion component, it cannot be compared directly to the baseline, although we expect that if it was combined with the baseline without query expansion then it may lead to an improvement.

Acknowledgments.

This work is supported by Science Foundation Ireland under grant 07/CE/I114.

References

[1] K. S. Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. In *Information Processing and Management*, pages 779–840, 2000.

[2] S. E. Robertson. On term selection for query expansion. *J. Doc.*, 46:359–364, December 1990.