Overview of the TREC-2012 Microblog Track

Ian Soboroff[1], Iadh Ounis[2], Craig Macdonald[2], Jimmy Lin[3,4*]
[1] NIST, Gaithersburg, MD, USA
[2] University of Glasgow, Glasgow, UK
[3] Twitter, San Francisco, CA, USA [4] University of Maryland, College Park, MD, USA
ian.soboroff@nist.gov, iadh.ounis@glasgow.ac.uk, craig.macdonald@glasgow.ac.uk, jimmylin@umd.edu

1. INTRODUCTION

The Microblog track examines search tasks and evaluation methodologies for information seeking behaviours in microblogging environments such as Twitter. It was first introduced in 2011, addressing a real-time adhoc search task, whereby the user wishes to see the most recent relevant information to the query. In 2012, the real-time adhoc task was changed slightly, and a new filtering task was added. The filtering task models a standing query where the user wishes to see relevant tweets as they are posted.

For the second year of the track, we reused the Tweets2011 corpus, described below. The corpus is comprised of 16M tweets distributed over two weeks, sampled courtesy of Twitter. The corpus was designed to be a reusable, representative sample of the twittersphere – i.e., both important and spam tweets were included. As the reusability of a test collection is paramount in TREC, we designed the corpus to be obtainable at any time by a researcher interested in conducting experiments. To accomplish this, in 2011, the TREC Microblog track introduced a novel methodology whereby participants sign an agreement for the ids of the tweets in the corpus. Tools are provided that permit the participants to download the corpus directly from Twitter.

The first Microblog track in TREC 2011 [3] was a remarkable success. In 2012, 40 groups participated in the track, with 33 groups submitting a total of 121 runs for the real-time adhoc task, and 19 groups submitting a total of 60 runs for the filtering task.

2. TWEETS2011 CORPUS

The TREC Microblog track in 2012 used the Tweets2011 corpus, which was created as part of last year's track. The corpus consists of an approximately 1% sample (after some spam removal) of tweets from January 23, 2011 to February 7, 2011 (inclusive), totaling approximately 16 million tweets. We summarize the corpus collection efforts here, but refer the reader to last year's track overview [3] for more details.

Creating a sharable reference collection of tweets is difficult because Twitter's terms of service forbid the redistribution of tweets. Last year, we devised a novel methodology whereby participants obtain a list of identifiers pointing to the tweets in the corpus after signing a usage agreement. Each identifier can be mapped to a URL at twitter.com which, when resolved, contains the tweet, delivered by Twitter according to their terms of service. Along with the corpus identifiers, we developed a set of tools to download a copy of the corpus, as well as sample code for indexing and retrieval.[1]

Note that individual downloads of the corpus would not be identical because tweets may have been deleted or made private since the corpus creation, and also some tweets may be unavailable due to transitory network failures. However, current evidence shows that corpus variability did not impact the evaluation results in the 2011 Microblog track [2]. We continue to monitor the availability and reusability of the collection.

3. REAL-TIME ADHOC TASK

In this section, we describe the real-time search task (Section 3.1), the pooling and judging procedures used (Section 3.2), and provide a brief overview of the results (Section 3.3).

3.1 Task Definition

A central aspect of search in microblog feeds is timeliness. In the *real-time search task*, we consider a user that makes a search query at a specific time, and wishes to see the most relevant information available up to that time. The real-time search task was the central task in 2011 [3] and this year underwent minor changes.

The main task for the 2012 Microblog track was the real-time search task, same as last year, where the user wishes to see the most recent and relevant information to the query. The real-time search task can be summarised as follows: At time t, find tweets about topic X [5]. This task is similar to adhoc search on Twitter's site, where a user's information need is represented by a query at a specific time [1].

For 2012, NIST created 60 new topics representing information needs at specific points in time. Figure 1 shows an example topic. The <querytime> tag contains the timestamp of the query in a human and machine readable ISO standard form, while the <querytweettime> tag contains the timestamp of the query in terms of the chronologically nearest tweet id in the corpus. While no narrative and description tags were provided to the participants, the topic developer recorded a clearly-defined information need for later use during assessment.

In last year's version of the task, participants were asked to rank relevant tweets by time. However, as reported in last year's track overview paper [3], this created significant ambiguity regarding how to interpret participants' retrieval scores and ranks. Moreover, the real-time nature of the task was not addressed within the run format, because the chronological order of tweets is invariant for any run. Indeed, real-time search doesn't necessarily mean that search results must be ranked (reverse) chronologically, rather, the only requirement is that an information need arrives at a specific time and concerns something happening right now.

*Certain companies and/or products may be identified in this paper in order to describe concepts and to specify experimental procedures adequately. Such identification is not intended to imply recommendation or endorsement by the National Institue of Standards and Technology, nor is it intended to imply that the companies or products identified are necessarily the best available for the purpose.

[1] http://twittertools.cc/

```
<top>
<num> Number: MB051 </num>
<query> British Government cuts </query>
<querytime> Tue Feb 08 23:56:46 +0000 2011 </querytime>
<querytweettime> 35124912364457984 </querytweettime>
</top>
```

Figure 1: Topic MB051 from the TREC 2012 Microblog track, real-time search task.

This year, we revised the task by asking participants to return a score for all tweets in the collection before the query-tweet-time, and only that score. Any unscored tweet was assumed to have a score of negative infinity.

For assessing the tweets, the assessors judged the relevance of a tweet after reading it and also by following any URLs linked from the tweet. Tweets were judged on the basis of the defined information need using a three-point scale:

Not Relevant. The content of the tweet does not provide any useful information on the topic, or is written in a language other than English, or is a retweet (RT).

Relevant. The tweet provides some information on the topic, but it is not sufficiently informative.

Highly Relevant. A highly relevant tweet will either contain highly informative content, or link to highly informative content. It is the hope that systems will score highly relevant tweets higher than relevant ones.

All assessments were conducted by NIST assessors. All tweets were judged in isolation, without trying to determine novelty with respect to older tweets. The assessor judged retweets as not relevant unless the retweet added content (e.g., prior to the 'RT' string) and thus added possible value beyond the original tweet. Assessors marked tweets in languages other than English as not relevant. When a tweet contained a link, the assessor took the linked page's content into consideration when deciding the relevance of the tweet.

An end-user application resembling Twitter's current search interface might apply a threshold on the tweet retrieval score and only show tweets above some threshold in chronological order. Alternatively, a search engine might choose to display the top-scoring tweets in rank order (regardless of time). Effectiveness in these notional applications is modeled by the task metrics. The main measures for the task this year were the receiver operating characteristic (ROC) curve and precision at rank 30. The ROC curve shows precision versus fallout for every possible score threshold. Precision at 30 (P@30) provides a simple measure of search effectiveness on early result pages, whether ranked by time or score. We also report average precision as a robust measure for ranked retrieval. All measures are computed with "highly relevant" as the required level of relevance.

3.2 Pooling and Judging

Participating groups were permitted to submit up to four runs to the real-time adhoc search task. At least one compulsory automatic run that does not use any external or future source of evidence was also requested. For the purposes of the task, we defined external and future evidence as follows:

External Evidence: Evidence outside the Tweets2011 corpus – for instance, this encompasses other tweets or information from Twitter, as well as other corpora, e.g., Wikipedia or the web.

Future Evidence: Information that would not have been available to the system at the timestamp of the query. For example, *idf* scores computed using tweets not already posted at the timestamp of the query.

The participating groups were encouraged to rank their submitted runs by preference. In addition to one compulsory automatic, real-time run that uses no external resources, the participating groups were at liberty to submit manual (i.e., not automatic, involving a human in the loop at some point in the run), external (i.e. using external resources) and untimely (i.e., not adhering to the real-time constraint) runs, which could be useful to improve the quality of the test collection. TREC received 121 runs from 33 participating groups. All runs were pooled to depth 100, according to the retrieval scores indicated in each run. Simple retweets were removed from the pools (as they were deemed to be non-relevant). The tweets were clustered so that textually similar tweets could be judged consistently.

3.3 Results

Tables 1 and 2 show the evaluation scores as well as metadata for all submitted runs, ordered by average precision at rank 30. Figure 2 shows ROC curves for the different runs from each group.

4. FILTERING TASK

This section describes the filtering task (Section 4.1), the evaluation measures used (Section 4.2), and provides a brief overview of the results (Section 4.3).

4.1 Task Definition

This year, we introduced a new task, tweet filtering. The filtering task is exactly the reverse of the real-time search task: a query arrives at a defined point in time, and the system must filter the subsequent stream of messages to select tweets relevant to the information need.

The filtering task is modeled on the TREC 2002 adaptive filtering task [4], and used topics MB 1–50 from last year's Microblog track. No new relevance judgments were created. The topics were retagged to reflect the altered meaning of the fields (see Figure 3). The earliest known relevant tweet in each of the 2011 topics was relabeled as the query-tweet-time trigger, and the original adhoc trigger was given as the endpoint, since no relevant tweets would exist after the original adhoc trigger tweet.

Topics with numbers $1, 6, 11 \ldots 46$, i.e., $(n \bmod 5 = 1)$, were allotted for participants to use for training their systems. In the testing phase, systems were allowed access to the topic fields, including the trigger tweet. The systems processed the tweets from the query-tweet-time to the query-newest-time, one at a time, making a decision on whether or not to show the tweet to the user. If the system decided to show the tweet, it could access the tweet's relevance judgment (if any) as immediate relevance feedback, but not otherwise.

Systems returned the list of tweets processed, each with their retrieval score and a decision yes/no indicating whether the tweet was shown to the user. Since no new pools or relevance judgments were made, the task was completely run using last year's data.

4.2 Measures

The filtering task used the TREC filtering measures [4]. Set precision and recall were computed over all retrieved tweets. The $F_{\beta=0.5}$ measure was then computed. Van Rijsbergen's F-measure is a function of precision and recall; the parameter β controls the

Run	group	P@30	MAP	manual?	RT?	docs?	extern?
hitURLrun3	HIT_MTLAB	0.2701	0.2642	automatic	yes	yes	no
uwatgclrman	UWaterlooMDS	0.2559	0.2277	manual	yes	no	no
hitLRrun1	HIT_MTLAB	0.2446	0.2411	automatic	yes	no	no
ICTWDSERUN1	ICTNET	0.2384	0.2093	automatic	yes	no	no
kobeL2R	KobeU	0.2384	0.2081	automatic	yes	no	no
kobeMHC2	KobeU	0.2356	0.2137	manual	yes	no	no
hitDELMrun2	HIT_MTLAB	0.2350	0.2257	automatic	yes	no	no
hitQryFBrun4	HIT_MTLAB	0.2345	0.2302	automatic	yes	no	no
kobeMHC	KobeU	0.2339	0.2115	manual	yes	no	no
ICTWDSERUN2	ICTNET	0.2339	0.1981	automatic	yes	no	no
PKUICST4	PKUICST	0.2333	0.2263	automatic	yes	no	no
cmuPrfPhrENo	CMU_Callan	0.2333	0.2223	automatic	yes	yes	no
otM12ihe	ot	0.2328	0.2259	automatic	no	no	no
tsqe	KobeU	0.2311	0.2093	automatic	yes	no	no
cmuPrfPhrE	CMU_Callan	0.2305	0.2200	automatic	yes	yes	no
FASILKOM01	FASILKOMUI	0.2294	0.1915	automatic	yes	no	yes
cmuPrfPhr	CMU_Callan	0.2266	0.2178	automatic	yes	no	no
IBMLTRFuture	IBM	0.2254	0.2018	automatic	yes	no	yes
IBMLTR	IBM	0.2237	0.1932	automatic	yes	no	no
uogTrLsE	uogTr	0.2232	0.2116	automatic	yes	yes	no
FASILKOM04	FASILKOMUI	0.2192	0.1810	automatic	yes	no	no
otM12ih	ot	0.2186	0.2036	automatic	no	no	no
uiucGSLIS01	uiucGSLIS	0.2186	0.1829	automatic	no	no	yes
FASILKOM02	FASILKOMUI	0.2186	0.1796	automatic	yes	no	yes
PKUICST1	PKUICST	0.2164	0.1639	automatic	yes	yes	no
FASILKOM03	FASILKOMUI	0.2153	0.1868	automatic	yes	no	yes
ICTWDSERUN3	ICTNET	0.2113	0.1878	automatic	yes	no	yes
PKUICST3	PKUICST	0.2113	0.1686	automatic	yes	yes	no
ICTWDSERUN4	ICTNET	0.2113	0.1650	automatic	yes	no	yes
otM12h	ot	0.2107	0.1911	automatic	yes	no	no
uwatrrfall	UWaterlooMDS	0.2107	0.1904	automatic	yes	no	no
york12mb3	york	0.2102	0.2009	automatic	yes	no	no
YORK1	york	0.2090	0.1907	automatic	yes	no	no
PKUICST2	PKUICST	0.2068	0.1561	automatic	yes	yes	no
uogTrBsE	uogTr	0.2062	0.1988	automatic	yes	yes	no
IBMBaseline	IBM	0.2028	0.1968	automatic	yes	no	no
prisRun4	BUPT_WILDCAT	0.2028	0.1555	automatic	yes	yes	no
XMRUN3	XMU_PANCHAO	0.2023	0.1802	automatic	yes	no	no
prisRun2	BUPT_WILDCAT	0.2017	0.1553	automatic	yes	no	no
otM12i	ot	0.2011	0.1755	automatic	no	no	no
KLIMLPLL	FUB	0.2006	0.1838	automatic	yes	no	yes
YORK2	york	0.2006	0.1694	automatic	yes	no	no
prisRun1	BUPT_WILDCAT	0.2006	0.1569	automatic	yes	no	no
AIrun1	AI_ROMA3	0.1994	0.1522	automatic	yes	no	yes
uiucGSLIS03	uiucGSLIS	0.1983	0.1717	automatic	yes	no	no
IRITbnetK	IRIT	0.1983	0.1715	automatic	yes	no	no
uwatrrflm	UWaterlooMDS	0.1977	0.1742	automatic	yes	no	no
uiucGSLIS02	uiucGSLIS	0.1972	0.1751	automatic	yes	no	no
cmuPhrE	CMU_Callan	0.1966	0.1854	automatic	yes	yes	no
york12mb4	york	0.1966	0.1694	automatic	yes	no	no
indri	udel	0.1960	0.1953	automatic	no	no	no
IRITbnetKSO	IRIT	0.1960	0.1717	automatic	yes	no	no
uwcmb12CP	waterloo	0.1955	0.1646	automatic	no	no	no
uwcmb12NT	waterloo	0.1949	0.1657	automatic	no	no	no
uwcmb12BL	waterloo	0.1944	0.1623	automatic	yes	no	no
UNCTQE	UNC_SILS	0.1938	0.1641	automatic	yes	no	no
KLIMLL	FUB	0.1932	0.1836	automatic	yes	no	no
XMRUN4	XMU_PANCHAO	0.1932	0.1575	automatic	yes	no	yes
KLIMLP1	FUB	0.1921	0.1949	automatic	yes	no	yes
QEWebFB	QCRI	0.1921	0.1710	automatic	yes	no	yes
prisRun3	BUPT_WILDCAT	0.1904	0.1453	automatic	yes	yes	no
IBCN2	UGENT_IBCN_SIS	0.1904	0.1408	automatic	yes	yes	no

Table 1: Adhoc runs, sorted by P@30 score, indicating run type (auto/manual), real-time (RT), if linked documents (docs?) and other external information were used (extern?). Continued on next page.

Table 2: Table 1, continued

Run	group	P@30	MAP	manual?	RT?	docs?	extern?
BM25PRF	qcri_twitsear	0.1898	0.1545	automatic	yes	no	no
QEWeb	QCRI	0.1881	0.1706	automatic	yes	no	yes
gucasQuery	GUCAS	0.1876	0.1503	automatic	yes	no	no
gucasGen	GUCAS	0.1876	0.1344	automatic	yes	no	no
KLIM	FUB	0.1870	0.1948	automatic	yes	no	no
mergedRun	qcri_twitsear	0.1864	0.1573	automatic	yes	no	no
uwcmb12CT	waterloo	0.1831	0.1620	automatic	yes	no	no
IBCN3	UGENT_IBCN_SIS	0.1825	0.1399	automatic	yes	yes	no
UNCRQE	UNC_SILS	0.1819	0.1490	automatic	yes	no	no
urlContent	SCIAITeam	0.1808	0.1465	automatic	yes	yes	no
BM25TRF	qcri_twitsear	0.1802	0.1503	automatic	yes	no	no
UNCQE	UNC_SILS	0.1802	0.1461	automatic	yes	no	no
gucasGenReg	GUCAS	0.1785	0.1318	automatic	yes	no	no
UvAfilter	UvA	0.1774	0.1385	automatic	yes	no	no
gucasBasic	GUCAS	0.1763	0.1476	automatic	yes	no	no
BAUjskls	BAU	0.1740	0.1527	automatic	yes	no	no
XMRUN1	XMU_PANCHAO	0.1723	0.1488	automatic	yes	no	no
XMRUN2	XMU_PANCHAO	0.1723	0.1487	automatic	yes	no	no
BLFB	QCRI	0.1718	0.1638	automatic	yes	no	no
BAUdfree	BAU	0.1718	0.1518	automatic	yes	no	no
BAUdph	BAU	0.1718	0.1510	automatic	yes	no	no
BL	QCRI	0.1701	0.1512	automatic	yes	no	no
csiroQEll112	csiro	0.1616	0.1393	automatic	no	no	no
UDInfoMBEx	udel_fang	0.1616	0.1161	automatic	no	no	yes
csiroNE112	csiro	0.1605	0.1363	automatic	no	no	no
BM25	qcri_twitsear	0.1605	0.1325	automatic	yes	no	no
BAUtf	BAU	0.1605	0.1320	automatic	yes	no	no
uiucGSLIS04	uiucGSLIS	0.1599	0.1259	automatic	yes	no	no
RUN3	uog_tw	0.1582	0.1297	automatic	yes	no	no
UNCTP	UNC_SILS	0.1559	0.1255	automatic	yes	no	no
csiroR112	csiro	0.1542	0.1324	automatic	yes	no	no
csiroQEt112	csiro	0.1537	0.1445	automatic	no	no	no
timemexp	udel	0.1531	0.0987	automatic	yes	no	no
IIEIR01	IIEIR	0.1508	0.1127	automatic	no	no	no
IIEIR03	IIEIR	0.1508	0.1088	automatic	no	no	no
IIEIR04	IIEIR	0.1508	0.1088	automatic	no	no	no
IIEIR02	IIEIR	0.1497	0.1073	automatic	no	no	no
UDInfoMBCW	udel_fang	0.1492	0.1161	automatic	no	no	yes
UDInfoMBIDF	udel_fang	0.1475	0.1040	automatic	yes	no	no
IBCN1	UGENT_IBCN_SIS	0.1469	0.1096	automatic	yes	no	no
baseline	SCIAITeam	0.1390	0.1224	automatic	yes	no	no
IRITfdvsmurl	IRIT	0.1390	0.0975	automatic	yes	yes	no
IBCN4	UGENT_IBCN_SIS	0.1379	0.1190	automatic	yes	yes	no
UDInfoMBTp	udel_fang	0.1373	0.0960	automatic	yes	no	no
aWekaModel	SCIAITeam	0.1350	0.1211	automatic	yes	no	yes
RUN2	uog_tw	0.1333	0.1089	automatic	yes	no	no
IRSIISI	IRSI	0.1333	0.0544	automatic	no	no	no
IRITfdvsm	IRIT	0.1311	0.0886	automatic	yes	no	no
IRSIISI1	IRSI	0.1311	0.0592	automatic	no	no	no
expansion	SCIAITeam	0.1282	0.0916	automatic	yes	no	yes
IRSIISI2	IRSI	0.1282	0.0508	automatic	no	no	no
IRSIISI3	IRSI	0.1260	0.0500	automatic	no	no	no
exttempwsf	UEdinburgh	0.1226	0.0700	automatic	yes	no	yes
RUN1	uog_tw	0.1192	0.0978	automatic	yes	no	no
exttempws	UEdinburgh	0.1192	0.0812	automatic	yes	no	no
timemodel	udel	0.1169	0.0694	automatic	yes	no	no
langluc	udel	0.1130	0.0647	automatic	yes	no	no
uogTrCIDE	uogTr	0.1124	0.0964	automatic	yes	yes	no
uwatgclrbase	UWaterlooMDS	0.0994	0.0777	automatic	yes	no	no

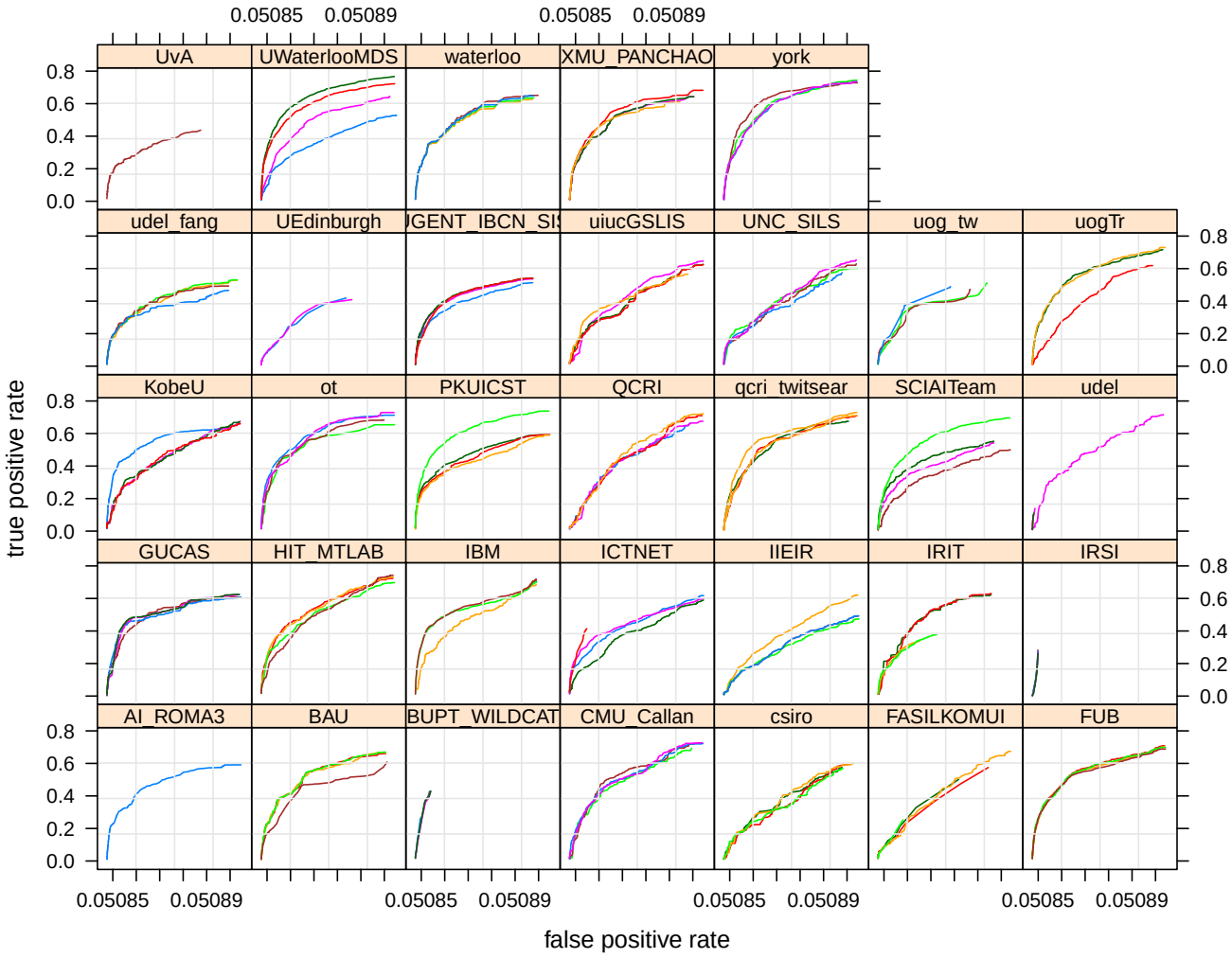

Figure 2: ROC curves for runs within each group.

```
<top>
<num> Number: MB002 </num>
<title> 2022 FIFA soccer  </title>
<querytime> Tue Feb 08 18:51:44 +0000 2011 </querytime>
<querytweettime> 29058771531595776 </querytweettime>
<querynewesttweet> 35048150574039040 </querynewesttweet>
</top>
```

Figure 3: Topic MB002 from TREC 2011 Microblog track, retagged for use in the 2012 filtering task.

relative weighting of each component:

$$F_\beta = \frac{(1+\beta^2)PR}{\beta^2 P + R}$$

A setting of $\beta = 0.5$ gives an emphasis to precision.

The other measure adopted from TREC filtering is linear utility. Imagine that a system receives a reward of two points for every relevant tweet retrieved, but takes a penalty of one point for every non-relevant tweet retrieved. Utility is the total points scored:

$$\textbf{T11U} = 2 \times \textit{relevant retrieved} - \textit{nonrelevant retrieved}$$

Filtering according to a linear utility function is equivalent to filtering by estimated probability of relevance, in this case, to retrieve if $P(rel) > 1/3$.

Utility values are unbounded, and hence need to be scaled to enable comparisons across topics. The utility scores are normalized to a fraction of their theoretical maximum, and scaled against an arbitrary minimum normalized utility value of -0.5 so that they may be averaged across topics:

$$\begin{aligned}
\textbf{MaxU} &= 2 \times \textit{total relevant} \\
\textbf{MinU} &= -0.5 \\
\textbf{NormU} &= \text{T11U}/\text{MaxU} \\
\textbf{T11SU} &= \frac{\max(\text{NormU}, \text{MinU}) - \text{MinU}}{1 - \text{MinU}}
\end{aligned}$$

Note that a T11SU value of $1/3$ can be achieved by a run that retrieves nothing — not helping but not wasting the user's time with non-relevant information either. This is called the "zero effort" baseline. TREC 2012 received 60 runs from 19 groups for the filtering task.

4.3 Results

Table 3 provides the evaluation results for each run, along with metadata included at submission time. Runs are sorted by descending T11SU score. Utility and F-measure are not always correlated, as seen in Figure 4. The scatterplot is shown with a diagonal line which would represent equal scores and a vertical line at the utility point of zero effort. While most "useful" runs with utility scores $> 1/3$ also have high F-measures, there is a wide range of F-measures that correspond to the zero-effort utility point.

5. CONCLUSIONS

This year marked the second iteration of the Microblog track, which featured a refinement of the real-time adhoc task and a new adaptive filtering task over tweets. The track continues to generate considerable interest, which we hope to sustain next year.

6. REFERENCES

[1] R.L.T. Santos, C. Macdonald, R. McCreadie, I. Ounis, and I. Soboroff. Information retrieval on the blogosphere. Foundations and Trends in Information Retrieval, 6:1, 2012, pp 1-125.

[2] R. McCreadie, I. Soboroff, J. Lin, C. Macdonald, I. Ounis, D. McCullough. On building a reusable Twitter corpus. *Proceedings of SIGIR 2012*.

[3] I. Ounis, C. Macdonald, J. Lin and I. Soboroff. Overview of the TREC-2011 Microblog track. In *Proceedings of TREC 2011*.

[4] S. Robertson, I. Soboroff. The TREC 2002 filtering track report. In *Proceedings TREC 2002*.

[5] I. Soboroff, D. McCullough, J. Lin, C. Macdonald, I. Ounis, R. McCreadie Evaluating real-time search over tweets. AAAI International Conference on Weblogs and Social Media (ICWSM 2012)

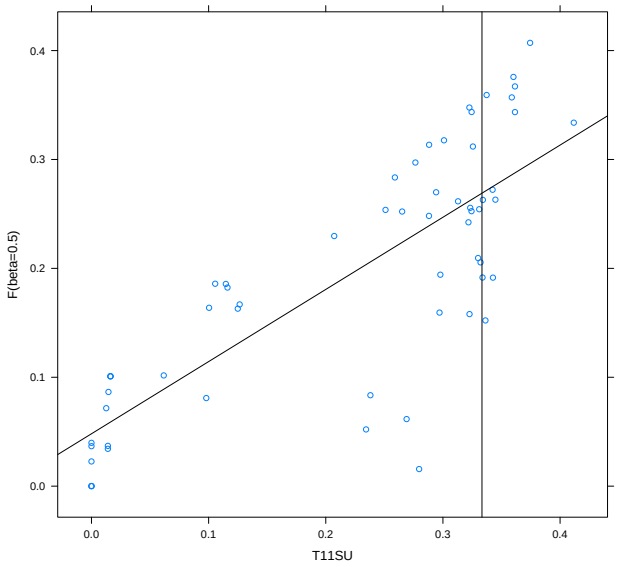

Figure 4: F_β versus scaled utility. The diagonal line shows where scores would lie if the two were equal. The vertical line shows the utility point of zero effort.

Run	group	Prec	Recl	F(0.5)	T11SU	manual?	RT?	docs?	extern?
hitUWT	HIT_MTLAB	0.6219	0.1740	0.3338	0.4117	automatic	yes	yes	no
PRISrun3	PRIS	0.4096	0.5675	0.4071	0.3744	automatic	yes	yes	no
PRISrun1	PRIS	0.3445	0.5981	0.3671	0.3615	automatic	yes	no	no
uogTrFADmN	uogTr	0.4206	0.3370	0.3435	0.3615	automatic	yes	no	no
PRISrun2	PRIS	0.3551	0.6273	0.3758	0.3602	automatic	yes	yes	no
expansion2	SCIAITeam	0.4405	0.3113	0.3570	0.3588	automatic	yes	no	no
csiroQERF111	csiro	0.4781	0.1427	0.2631	0.3448	automatic	yes	no	no
uogTrFFDmN	uogTr	0.5508	0.1394	0.1915	0.3427	automatic	yes	no	no
PKUICSTF1	PKUICST	0.3963	0.2300	0.2722	0.3424	automatic	yes	no	no
expansionurl	SCIAITeam	0.4282	0.3448	0.3592	0.3373	automatic	yes	yes	no
udellngth	udel	0.3381	0.0959	0.1522	0.3363	automatic	yes	no	no
PKUICSTF4	PKUICST	0.3766	0.2936	0.2629	0.3341	automatic	yes	yes	no
okapiv1	unir_de	0.3370	0.1024	0.1916	0.3338	automatic	yes	no	no
window2run	HIT_MTLAB	0.3860	0.0987	0.2055	0.3321	automatic	yes	no	no
FasilkomF1	FASILKOMUI	0.2822	0.2596	0.2543	0.3310	automatic	yes	no	yes
uogTrFFeDm	uogTr	0.4940	0.1621	0.2095	0.3300	automatic	yes	no	no
weka	SCIAITeam	0.4261	0.2384	0.3119	0.3256	automatic	yes	no	yes
RetrievalThr	QCRI	0.3571	0.4651	0.3436	0.3245	automatic	yes	no	no
PKUICSTF2	PKUICST	0.3701	0.2809	0.2525	0.3244	automatic	yes	no	no
PKUICSTF3	PKUICST	0.3857	0.2272	0.2556	0.3233	automatic	yes	yes	no
csiroSVMqe111	csiro	0.1953	0.2217	0.1580	0.3227	automatic	yes	no	no
PRISrun4	PRIS	0.3194	0.6676	0.3477	0.3226	automatic	yes	no	no
FasilkomF3	FASILKOMUI	0.2645	0.2410	0.2423	0.3218	automatic	yes	no	no
FasilkomF2	FASILKOMUI	0.2850	0.2983	0.2616	0.3129	automatic	yes	no	yes
basic	SCIAITeam	0.3663	0.3158	0.3176	0.3009	automatic	yes	no	no
okapiv2rel	unir_de	0.2831	0.1486	0.1942	0.2978	automatic	yes	no	no
csiroshuq111	csiro	0.2688	0.2294	0.1594	0.2971	automatic	yes	no	no
hitRSW	HIT_MTLAB	0.2838	0.3440	0.2699	0.2942	automatic	yes	no	no
york12bd1i	york	0.3412	0.4317	0.3135	0.2882	automatic	yes	no	no
udelcosrun	udel	0.3407	0.2533	0.2482	0.2881	automatic	yes	no	no
irsicombsum	IRSI	0.0523	0.0049	0.0157	0.2797	automatic	no	no	no
UnifiedThr	QCRI	0.3186	0.4225	0.2972	0.2765	automatic	yes	no	no
vsmv1	unir_de	0.1217	0.0732	0.0616	0.2690	automatic	yes	no	no
FasilkomF4	FASILKOMUI	0.3011	0.2471	0.2522	0.2652	automatic	yes	no	yes
FRUN3	uog_tw	0.3414	0.4440	0.2835	0.2590	automatic	yes	no	no
uogTrFADmI	uogTr	0.3197	0.3868	0.2537	0.2510	manual	yes	no	no
vsmv2rel	unir_de	0.1411	0.0518	0.0835	0.2381	automatic	yes	no	no
irsivoting	IRSI	0.0939	0.0327	0.0521	0.2344	automatic	no	no	no
FRUN1	uog_tw	0.2657	0.4238	0.2297	0.2072	automatic	yes	no	no
ICTNETFTRUN1	ICTNET	0.1553	0.5020	0.1669	0.1265	automatic	yes	no	no
ICTNETFTRUN2	ICTNET	0.1513	0.5244	0.1630	0.1249	automatic	yes	no	yes
BAUdfreef	BAU	0.1681	0.6021	0.1824	0.1161	automatic	yes	no	no
BAUjsklsf	BAU	0.1708	0.6023	0.1857	0.1147	automatic	yes	no	no
BAUdphf	BAU	0.1712	0.5972	0.1859	0.1056	automatic	yes	no	no
BAUdfi0f	BAU	0.1495	0.5932	0.1638	0.1004	automatic	yes	no	no
csirolrhuq111	csiro	0.0821	0.3431	0.0809	0.0980	automatic	yes	no	no
FRUN2	uog_tw	0.1099	0.4928	0.1017	0.0617	automatic	yes	no	no
gucasB	GUCAS	0.0848	0.6656	0.1009	0.0162	automatic	yes	no	no
gucasL1	GUCAS	0.0848	0.6656	0.1009	0.0162	automatic	yes	no	no
gucasL2	GUCAS	0.0848	0.6656	0.1009	0.0162	automatic	yes	no	no
uw	UWaterlooMDS	0.0740	0.4578	0.0865	0.0145	automatic	yes	no	no
nemisExt	NEMIS_ISTI_CNR	0.0293	0.4433	0.0343	0.0140	automatic	yes	yes	yes
nemisNotExt	NEMIS_ISTI_CNR	0.0315	0.4232	0.0370	0.0140	automatic	yes	no	no
QFilRun3	qcri_twitsear	0.0617	0.6096	0.0716	0.0126	automatic	yes	no	no
ICTNETFTRUN3	ICTNET	0.0000	0.3641	0.0000	0.0000	automatic	yes	no	no
ICTNETFTRUN4	ICTNET	0.0001	0.4933	0.0001	0.0000	automatic	yes	no	yes
QFilRun1	qcri_twitsear	0.0305	0.7256	0.0367	0.0000	automatic	no	no	no
QFilRun2	qcri_twitsear	0.0331	0.7226	0.0398	0.0000	automatic	yes	no	no
urlAllFB	HIT_MTLAB	0.0000	0.9146	0.0001	0.0000	automatic	yes	yes	no
uwn	UWaterlooMDS	0.0184	0.8274	0.0227	0.0000	automatic	yes	no	no

Table 3: Filtering runs, sorted by T11SU score, indicating the same metadata as shown for adhoc runs in Table 1.