Overview of the TREC-2011 Microblog Track

Iadh Ounis[1], Craig Macdonald[1], Jimmy Lin[2,3], Ian Soboroff[4*]

[1] University of Glasgow, Glasgow, UK
[2] Twitter, San Francisco, CA, USA [2] University of Maryland, College Park, MD, USA
[4] NIST, Gaithersburg, MD, USA

first.lastname@glasgow.ac.uk, jimmylin@umd.edu, ian.soboroff@nist.gov

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 but relevant information to the query. In particular, systems should respond to a query by providing a list of relevant tweets ordered from newest to oldest, starting from the time the query was issued.

For TREC 2011, we used the newly-created Tweets2011 corpus. The corpus is comprised of 16M tweets over approximately two weeks, sampled courtesy of Twitter. The corpus is designed to be a reusable, representative sample of the twittersphere. As the reusability of a test collection is paramount in a TREC track, these sampled tweets can be obtained at any point in time (subjected to some caveats, discussed below). To accomplish this, the TREC Microblog track introduced a novel methodology whereby participants sign an agreement for the ids of the tweets in the corpus. Tools are then provided that permit the downloading of the corpus from the Twitter website.

The first Microblog track in TREC 2011 has been a remarkable success. A total of 59 groups participated in the track from across the world, with 184 submitted runs.

2. TWEETS2011 CORPUS

The corpus was obtained using a donation of the unique identifiers of a sample of tweets from Twitter. Creating a sharable reference collection of tweets is difficult, because Twitter's terms of service forbids the redistribution of tweets. Among other reasons for this, Twitter users can delete their tweets (and indeed their entire account) or restrict their tweets to followers only, and these states can change during and outside the corpus epoch. We devised a novel methodology whereby participants obtain a list of identifiers pointing to the tweets in the corpus after signing a usage agreement. These identifiers are of the form (`screen_name`, `tweet_id`). 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.

We developed a set of tools to generate a copy of the corpus given the list of tweet ids, as well as sample indexing and searching code.[1] Participants and others obtaining the Tweets2011 collection agree not to redistribute the tweets in the collection, but anyone can obtain a substantially identical tweet set using the ids and tools. The set is only "substantially" identical because tweets may have been deleted or made private in the intervening time, and also some tweets may be unavailable due to transitory network failures.

The resulting corpus, called Tweets2011, 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. Major events that took place within this time frame include the massive democracy demonstrations in Egypt as well as the Super Bowl in the United States. Each day of the corpus is split into files called *blocks*, each of which contains about 10,000 tweets compressed using gzip. Each tweet is in JSON format, similar (but not identical) to the format used by the Twitter streaming hoses. Within the corpus, tweets are ordered by tweet ids, which are roughly chronologically ordered for our purposes. The sample of tweets and the corresponding tools were released to the TREC participating groups on 16th May 2011.

3. REAL-TIME SEARCH TASK

3.1 Task Definition

In TREC 2011, the Microblog track addressed one single pilot task, entitled the *real-time search task*, where the user wishes to see the most recent but relevant information to the query. The real-time search task can be summarised as: *At time t, find tweets about topic X*. This task is akin to adhoc search on Twitter, where a user's information need is represented by a query at a specific time. Participants were asked to rank the relevant tweets by time. One possible interpretation of the task is to rank all tweets up to time t, keep all *interesting* tweets, and then discard non-relevant tweets. Interestingness is subjective, but the issuer of a query might interpret it as providing added value with respect to the query topic. It is of note that for TREC 2011, the novelty between tweets was not considered.

NIST created 50 new topics based on the Tweets2011 collection, each representing an information need at a specific point 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 within the corpus. Moreover, while no narrative and description tags were provided to participants during the evaluation (as with earlier TREC adhoc topics), the topic developer created a clearly defined information need for later use during assessment.

*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/, which redirected at the time of writing to https://github.com/lintool/twitter-tools/

```
<top>
<num> Number: MB01 </num>
<title> Wael Ghonim </title>
<querytime> 25th February 2011 04:00:00 +0000 </querytime>
<querytweettime> 3857291841983981 </querytweettime>
</top>
```

Figure 1: Topic MB01 from the TREC 2011 Microblog track.

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 either written in a language other than English, or is a retweet.

Relevant. The tweet mentions or provides some minimally useful information on the topic.

Highly Relevant. A highly relevant tweet will either contain highly informative content, or link to highly informative content.

All assessments were conducted by NIST assessors. The primary evaluation measure was precision at rank 30 cutoff.

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 sources of evidence was also requested. For the purposes of the task, we defined external and future evidence as follows:

External Evidence: Evidence beyond 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. For comparison purposes, the track requested at least one compulsory automatic run that abides by real-time and external resource constraints; beyond this, the participating groups were at liberty to submit manual, external and untimely runs, which could be useful to improve the quality of the test collection. TREC received 184 runs in total, from 59 participating groups. All runs were pooled to depth 30, according to the ranking indicated in each run. We later determined that this pooling process was problematic, but the problems did not affect the evaluation results reported here. We elaborate on the problems in the Discussion section below.

Simple retweets were removed from the pools (as they were *de facto* assumed to be non-relevant). Tweets were clustered to bring near-duplicates close together in the pools, using shingling [1].[2] We believe this sorting supported consistent judgments because the assessor would judge lexicographically similar tweets together, but we did not measure the effect on assessor consistency.

[2]Usually in TREC, pools are sorted by document identifier. The goal in pool sorting is to sort without respect to run, score, or relevance.

Measure	All Relevant		Highly Relevant	
	Best	Median	Best	Median
P@30	0.6116	0.2575	0.2646	0.0687
MAP	0.5127	0.1426	0.4740	0.1377

Table 1: Summary of results from the TREC 2011 Microblog track evaluation: per-topic best and medians for the 49 topics where all relevant and highly relevant tweets were considered relevant (denoted All Relevant), and the 33 topics where only highly relevant tweets were considered relevant (denoted Highly Relevant).

3.3 Results

We first report evaluation results of the 59 participating groups with 49 topics. Topic 50 did not have any relevant tweets in the pool, and it was therefore dropped from the evaluation. As mentioned in Section 3.1, the primary measure for retrieval effectiveness was precision at rank 30 (P@30), but we also report mean average precision (MAP). Table 1 shows the per-topic best and medians of the submitted real-time search task runs. Since only 33 topics have tweets judged to be highly relevant in the pool, the table shows two separate sets of scores. The first considers all relevant and highly relevant tweets as relevant and is over 49 topics. The second considers only highly relevant tweets, and is over 33 topics. From the results, it appears that the real-time search task is challenging when we focus on only the highly-relevant tweets.

In the next analysis, we focus on the evaluation results using all 49 topics, where all relevant and highly relevant tweets are considered as relevant. Table 2 shows the best submitted compulsory runs from each participating group, ranked by P@30. Although this condition was required, not all groups followed the requirement; the 14 groups which did not submit compulsory runs are omitted. Mean Average Precision (MAP) is also reported in the table. The correlation between the ranking of groups by MAP and P@30 is high but not without noticable differences (Spearman's $\rho = 0.82$). Using a bootstrap test for discriminative power [5], differences in P@30 or MAP of less than 0.07 have a run swap probability of greater than 5%, and thus are not deemed to be meaningful.

Table 3 shows the best performing run from each participating group, regardless of the run type, and the extent to which it abides by the real-time and external resources constraints. In contrast to Table 2, all 59 groups are present. We note the presence of five manual runs. Yet, the overall ranking of groups is not markedly different with the relaxing of the run constraints – we observe a correlation of $\rho = 0.93$ between the ranking of groups by best compulsory run and best run (among those groups that submitted a compulsory run).

Finally, Table 4 shows the best submitted compulsory run from each participating group, ranked by P@30 calculated using only highly relevant tweets. This ranking of groups is nearly the same as when all relevant tweets are counted (table 2, $\rho = 0.95$), although in some cases a group's best run was not the same under the two conditions. Only 33 topics have highly relevant tweets, and differences of less than 0.03 P@30 and 0.08 MAP are not meaningful according to the bootstrap discriminative power test.

4. DISCUSSION

As mentioned above, tweets were pooled from participating runs down to rank 30, following the rank field of the run. This was problematic for two reasons. The first reason is that this is itself an unusual pooling approach for TREC; runs are traditionally pooled by the document score (retrieval status value, also known as the

Group	Run	Auto.	Corpus	Real-time	Linked	Ext. Res.	P@30	MAP
isi	isiFDL	✔	HTML	✔			0.4551	0.1892
FUB	DFReeKLIM30	✔	HTML	✔			0.4401	0.2316
PRIS	PRISrun1		HTML	✔			0.4388	0.3302
KobeU	ri	✔	HTML			✔	0.4265	0.2203
CLARITY_DCU	clarity1	✔	HTML	✔			0.4211	0.2109
FASILKOMUI	FASILKOM02	✔	HTML	✔		✔	0.4197	0.1904
waterloo	waterlooa3	✔	HTML	✔		✔	0.4095	0.2082
ICTIR	run2		HTML				0.4075	0.2953
Purdue_IR	myrun2	✔	HTML	✔			0.3993	0.1977
HIT_LTRC	hitWIt	✔	HTML	✔			0.3973	0.3157
wis_tudelft	WISTUD	manual	HTML				0.3946	0.2719
PKU_ICST	PKUICST2	✔	HTML	✔			0.3905	0.2196
CIIR	ciirRun2	✔	HTML				0.3646	0.2274
SEEM_CUHK	WiseFifthRun	✔	HTML	✔			0.3578	0.1687
NUSIS	relevanceRun	✔	HTML	✔			0.3517	0.1862
syles	sylesNoRes	✔	HTML	✔			0.3476	0.2114
KAUST	KAUSTRerank	✔	HTML	✔			0.3456	0.1699
DUTIR	dutirMixFb	✔	HTML	✔			0.3408	0.2902
IRSI	Google1GNO	✔	HTML			✔	0.3401	0.2265
gslisUIUC	gut	✔	HTML	✔			0.3218	0.1233
QCRI	QCRIwTagOrg	✔	HTML				0.3177	0.1230
RMIT	RMITMRR		HTML			✔	0.3163	0.2311
SienaCLTeam	SienaCL1B	✔	HTML		✔		0.3082	0.1635
udel	udelIndri	✔	HTML				0.3082	0.1230
COMMIT	COMMITlinks	✔	JSON	✔	✔	✔	0.3027	0.1703
Udel_Fang	UDMicroIDF	✔	HTML	✔			0.3027	0.1842
DLDE	omarRun	✔	HTML	✔			0.2932	0.0874
UPorto	baseline2	✔	JSON	✔			0.2925	0.1239
UIowaS	UIowaS3	✔	HTML		✔	✔	0.2918	0.1403
UoW	PL2NoQeSd	✔	HTML	✔			0.2823	0.1561
PolyU	LJQO5	✔	HTML	✔			0.2639	0.1633
xmuPRC	RunPure	✔	HTML	✔			0.2639	0.1145
IRIT_SIG	iritfd1	✔	HTML	✔		✔	0.2605	0.2115
kwcenter	2	✔	HTML				0.2578	0.1905
UniMelbLT	melblt	✔	HTML	✔			0.2565	0.1409
UICIR	uicir1	✔	HTML		✔	✔	0.2524	0.0916
yandex	ya4	✔	other		✔	✔	0.2381	0.0822
L3S	qHtagBaseRun	✔	HTML				0.2190	0.1154
QUT1	run3a	✔	other	✔			0.2034	0.0663
uogTr	uogTrUB2	✔	HTML	✔			0.1939	0.1014
WeST	WESTfilext	✔	HTML	✔		✔	0.1776	0.1071
Vitalie_Scurtu	scurtuRun1	✔	HTML	✔			0.1762	0.1453
NEMIS_ISTI_CNR	runNeMISext	✔	HTML		✔	✔	0.1714	0.1186
FDUMED	FDUNLP	✔	HTML	✔			0.1510	0.1411
Elly	Basic	✔	HTML	✔			0.1463	0.0943
SIEL_IIITH	sielrun4	✔	HTML	✔			0.1265	0.0569
GUCAS	IDEAACTQE	✔	HTML	✔			0.1190	0.1106
Morpheus	MorpheusRun1	✔	HTML	✔			0.1150	0.0206
UGLA_D	tfTP01	✔	JSON	✔			0.1007	0.1166
UCSC	run3	✔	HTML		✔		0.0939	0.1416
monash	MONASH1NEW	✔	HTML	✔			0.0823	0.1144
ikm101	ikmRun1		HTML		✔		0.0612	0.0433
ICTNET	ICTNET11MBR3	✔	HTML	✔		✔	0.0490	0.1000
TUD_DMIR	EMAX	✔	JSON	✔			0.0435	0.0301
ULuga	baselineBM25	✔	HTML	✔			0.0415	0.0292
KapeReunion	kapeRun	✔	HTML	✔			0.0401	0.0553
utwente	UTWngFuture	✔	other				0.0245	0.0246
uiuc	uiucsf	✔	HTML				0.0075	0.0007

Table 3: Ranked runs, 1 per group; ranked by P@30 where tweets judged highly or minimally relevant are considered relevant.

Group	Run	P@30	MAP
isi	isiFDL	0.4551	0.1892
FUB	DFReeKLIM30	0.4401	0.2316
PRIS	PRISrun1	0.4388	0.3302
CLARITY_DCU	clarity1	0.4211	0.2109
FASILKOMUI	FASILKOM01	0.4184	0.1809
Purdue_IR	myrun2	0.3993	0.1977
ICTIR	run1fix	0.3986	0.2444
HIT_LTRC	hitWIt	0.3973	0.3157
PKU_ICST	PKUICST2	0.3905	0.2196
waterloo	waterlooa4	0.3755	0.1871
SEEM_CUHK	WiseFifthRun	0.3578	0.1687
NUSIS	relevanceRun	0.3517	0.1862
syles	sylesNoRes	0.3476	0.2114
KAUST	KAUSTRerank	0.3456	0.1699
CIIR	ciirRun1	0.3449	0.2005
DUTIR	dutirMixFb	0.3408	0.2902
gslisUIUC	gut	0.3218	0.1233
KobeU	rmal	0.3136	0.1594
Udel_Fang	UDMicroIDF	0.3027	0.1842
SienaCLTeam	SienaCLbase	0.2939	0.1498
DLDE	omarRun	0.2932	0.0874
UPorto	baseline2	0.2925	0.1239
UoW	PL2NoQeSd	0.2823	0.1561
PolyU	LJQO5	0.2639	0.1633
xmuPRC	RunPure	0.2639	0.1145
COMMIT	COMMITbase	0.2585	0.2026
kwcenter	3	0.2578	0.1905
IRIT_SIG	iritfd2	0.2565	0.1920
UniMelbLT	melblt	0.2565	0.1409
yandex	YNDXTPC2	0.2156	0.1026
QUT1	run3a	0.2034	0.0663
uogTr	uogTrUB2	0.1939	0.1014
Vitalie_Scurtu	scurtuRun1	0.1762	0.1453
WeST	WESTfilter	0.1680	0.1109
FDUMED	FDUNLP	0.1510	0.1411
Elly	Basic	0.1463	0.0943
SIEL_IIITH	sielrun4	0.1265	0.0569
GUCAS	IDEAACTQE	0.1190	0.1106
Morpheus	MorpheusRun1	0.1150	0.0206
UGLA_D	tfTP01	0.1007	0.1166
wis_tudelft	basicWISTUD	0.0993	0.1110
UCSC	cyfrun1	0.0932	0.1309
monash	MONASH1NEW	0.0823	0.1144
ICTNET	ICTNET11MBR1	0.0476	0.1039
TUD_DMIR	EMAX	0.0435	0.0301
ULuga	baselineBM25	0.0415	0.0292
KapeReunion	kapeRun	0.0401	0.0553
utwente	UTBase	0.0163	0.0103

Table 2: Automatic runs abiding by the real-time and external resources constraints, 1 per group; ranked by P@30, where tweets judged highly or minimally relevant are considered relevant.

Group	Run	P@30	MAP
PRIS	PRISrun2	0.1687	0.3135
isi	isiFDRM	0.1566	0.2476
FUB	DFReeKLIM30	0.1495	0.2286
CLARITY_DCU	clarity1	0.1434	0.2064
PKU_ICST	PKUICST2	0.1414	0.2380
HIT_LTRC	hitWIt	0.1354	0.2404
ICTIR	run1fix	0.1354	0.2352
KAUST	KAUSTRerank	0.1273	0.1201
Purdue_IR	myrun3	0.1253	0.1998
syles	sylesNoRes	0.1202	0.1902
CIIR	ciirRun1	0.1162	0.1935
DUTIR	dutirMixFb	0.1162	0.2351
SEEM_CUHK	WiseFouthRun	0.1152	0.1606
FASILKOMUI	FASILKOM01	0.1081	0.0971
Udel_Fang	UDMicroIDF	0.1081	0.2279
COMMIT	COMMITbase	0.1051	0.1930
waterloo	waterlooa4	0.1010	0.1608
IRIT_SIG	iritfd2	0.0960	0.1621
UPorto	baseline2	0.0949	0.0983
NUSIS	balanceRun	0.0939	0.1402
gslisUIUC	gut	0.0929	0.0833
PolyU	LJQO5	0.0899	0.1494
KobeU	rmal	0.0869	0.1582
UniMelbLT	melblt	0.0828	0.1579
UoW	PL2NoQeSd	0.0818	0.1608
uogTr	uogTrUB2	0.0818	0.0714
kwcenter	1	0.0808	0.1529
SienaCLTeam	SienaCLbase	0.0768	0.1329
xmuPRC	RunPure	0.0727	0.0516
DLDE	omarRun	0.0717	0.0476
yandex	YNDXTPC2	0.0697	0.1265
FDUMED	FDUNLP	0.0677	0.1707
Elly	Basic	0.0566	0.0871
QUT1	run3a	0.0556	0.0646
Vitalie_Scurtu	scurtuRun1	0.0535	0.1590
WeST	WESTfilter	0.0515	0.0887
GUCAS	IDEAACTQE	0.0434	0.1153
SIEL_IIITH	sielrun4	0.0394	0.0235
Morpheus	MorpheusRun1	0.0384	0.0130
UCSC	cyfrun1	0.0384	0.1501
UGLA_D	tfTP01	0.0374	0.1017
monash	MONASH1NEW	0.0323	0.1485
wis_tudelft	basicWISTUD	0.0323	0.1207
ICTNET	ICTNET11MBR1	0.0242	0.1444
KapeReunion	kapeRun	0.0172	0.0899
TUD_DMIR	RTB	0.0101	0.0035
ULuga	baselineBM25	0.0091	0.0141
utwente	UTBase	0.0010	0.0045

Table 4: Automatic runs abiding by the real-time and external resources constraints, 1 per group; ranked by P@30, where only tweets judged to be highly relevant are considered relevant.

‘sim’ field) with ties in score broken by document identifier (in other words, randomly with respect to relevance). The runs were pooled by rank following the premise that because the task concerned real-time retrieval, the rank order in the runs was significant.

In fact, this premise itself revealed the second problem, which is that the task was underspecified. Some participants computed scores to correspond to the order of tweet ids. Some participants adjusted ranks to boost documents, without changing scores. Some participants submitted “traditional” output with ranks coerced to increase with decreasing retrieval status values. In short, the task did not define the semantics of the run submission sufficiently for us to make comparisons between different systems. Comparisons of runs within a group are valid as long as the run ranking has the same meaning, but comparisons between systems in different groups are less valid without further investigation.

These issues could have two practical effects. One is that by pooling to a depth different than the depth semantics of the run, we may not have judged an equal number of tweets from each run. While pooling to equal depths is not necessary (see for example [2]), this is usually done in TREC in the spirit of fairness to all participants.

5. CONCLUSIONS

The Microblog track ran for the first time at TREC 2011, addressing a real-time adhoc search task. The creation of the corpus, which followed a novel methodology for TREC, has been a major success. With 59 groups participating, this is the largest TREC track/task ever in terms of participating groups. The evaluation results show that the real-time search task is far from being a solved problem. The Microblog track will run again in TREC 2012.

6. REFERENCES

[1] A.Z. Broder, S.C. Glassman, M.S. Manasse. Syntactic clustering of the Web. In *Proceedings of the 6th International World Wide Web Conference (WWW 1997),* Santa Clara, California, USA, 1997.

[2] G.V. Cormack, C.R. Palmer and C.L.A. Clarke. Efficient construction of large test collections. In *Proceedings of the 21st Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998),* Melbourne, Australia, 1998.

[3] C. Macdonald, R.L.T. Santos, I. Ounis and I. Soboroff. Blog track research at TREC. SIGIR Forum, 44(1):58–75, 2010.

[4] I. Ounis, M. de Rijke, C. Macdonald, G. Mishne and I. Soboroff. Overview of TREC-2006 Blog track. In *Proceedings of TREC 2006,* Gaithersburg, Maryland, USA, 2007.

[5] T. Sakai. Evaluating evaluation metrics based on the bootstrap. in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Seattle, Washington, USA, 2006.