Crowdsourcing Blog Track Top News Judgments at TREC

Richard McCreadie
School of Computing Science
University of Glasgow
Glasgow, G12 8QQ
richardm@dcs.gla.ac.uk

Craig Macdonald
School of Computing Science
University of Glasgow
Glasgow, G12 8QQ
craigm@dcs.gla.ac.uk

Iadh Ounis
School of Computing Science
University of Glasgow
Glasgow, G12 8QQ
ounis@dcs.gla.ac.uk

ABSTRACT

Since its inception, the venerable TREC retrieval conference has relied upon specialist assessors or participating groups to create relevance judgments for the tracks that it runs. However, recently crowdsourcing has been proposed as a possible alternative to traditional TREC-like assessments, supporting fast accumulation of judgments at a low cost. 2010 was the first year that TREC experimented with crowdsourcing. In this paper, we report our successful experience in creating relevance assessments for the TREC Blog track 2010 top news stories task. We conclude that crowdsourcing is an effective alternative to using specialist assessors or participating groups for this task.

1. INTRODUCTION

Relevance assessments is a crucial component when developing and evaluating information retrieval (IR) systems like search engines. Since its inception in 1992, the *T*ext *RE*trieval *C*onference (TREC) has played an important role in the IR community, creating reusable test collections and relevance assessments for a series of IR tasks. This has been underpinned by robust relevance assessments by specialist TREC assessors, or by the participating groups themselves.

However, this style of assessments also holds some profound limitations. Most notably, judgement by TREC assessors is expensive in terms of time and resources, while not being greatly scalable [2]. Furthermore, while engaging the participants for judging is free, the volume of judgments that can be produced is limited by the number of participants to the task in question.

On the other hand, *crowdsourcing* [9] has been championed as a viable method for creating relevance assessments, and indeed, as an alternative to traditional TREC assessments [2]. The reputed advantages of crowdsourcing are four-fold: judging can be performed quickly, cheaply, at a larger scale and with redundancy to achieve sufficient quality [3]. However, crowdsourcing has also been the subject of much controversy as to its effectiveness, in particular with regard to the lower quality of work produced [5], the lack of motivation for workers due to below-market wages [6] and susceptibility to malicious workers [7].

In TREC 2010, the Blog track examined real-time news story ranking within the blogosphere. In particular, participants were asked to rank news stories for a day of interest by their relative importance on that day, based upon evidence from the blogosphere [10]. Notably, 'importance' in this case is relative to the other stories published upon the same day. In this paper, we describe our successful experience when crowdsourcing relevance judgments for the Blog track top news stories task. Our contributions are three-fold: 1) we summarise the first successful instance of crowdsourcing at TREC, 2) we quantitatively assess both the crowdsourcing job itself, as well as the judgments produced and 3) we propose best practices based upon experience gained.

The structure of this paper is as follows. Section 2 describes the task that we crowdsourced, in addition to the interface and experimental setup employed. In Section 3, we detail the research questions that we investigate with regard to our crowdsourcing of relevance judgments for TREC 2010 and describe our experimental results. We provide concluding remarks in addition to some best practices in Section 4.

2. JUDGING NEWS STORY IMPORTANCE

The task that we address in this paper is the crowdsourcing of relevance assessments (qrels) for the Blog track top news stories task at TREC 2010. In particular, for each day of interest (query day), the participating systems returned a ranking of 50 news stories that they deemed important on that day for each of 5 news categories, namely: U.S., World, Sport, Business/Financial and Science/Technology news. The rankings from the participants were sampled using statMAP sampling [4], to a depth of 32 stories per day and category, resulting in 160 stories per day to be judged, with 8,000 stories in total [10]. The relevance assessment task is to label each of these sampled stories as important or not from an editorial perspective, such that a system's ranking based upon the blogosphere can be compared to that produced by a newspaper editor. In the following subsections we detail our crowdsourcing methodology as well as the interface that was used.

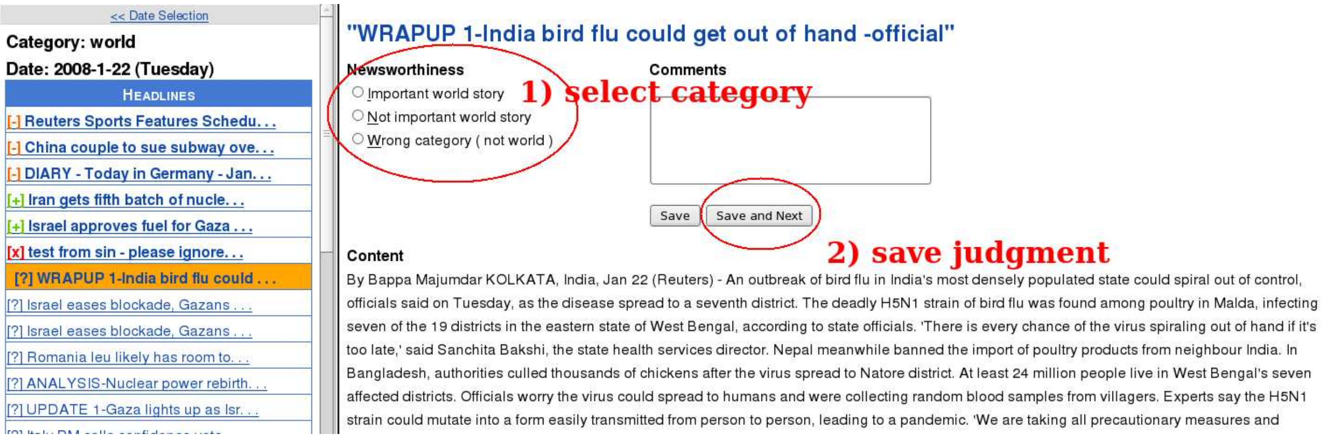
2.1 Crowdsourcing Task

We used Amazon's online marketplace Mechanical Turk (MTurk) to perform our judging. In particular, each MTurk Human Intelligence Task (HIT) covers the 32 top stories sampled for a single day and news category. For these stories, we ask workers to judge each as either: 1) Important and of the correct category, 2) Not important but of the correct category or 3) of the wrong category. To inform this judgement, the worker was presented with both the head-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WSDM 2011 Workshop on Crowdsourcing for Search and Data Mining (CSDM 2011), Hong Kong, China, Feb. 9, 2011

Copyright is held by the author/owner(s).



<< Date Selection
HEADLINES
[-] Reuters Sports Features Schedu. . .
[-] China couple to sue subway ove. . .
[-] DIARY - Today in Germany - Jan. . .
[+] Iran gets fifth batch of nucle. . .
[+] Israel approves fuel for Gaza . . .
[x] test from sin - please ignore. . .
[?] WRAPUP 1-India bird flu could . . .
[?] Israel eases blockade, Gazans . . .
[?] Israel eases blockade, Gazans . . .
[?] Romania leu likely has room to. . .
[?] ANALYSIS-Nuclear power rebirth. . .
[?] UPDATE 1-Gaza lights up as Isr. . .
Newsworthiness
○ Important world story
○ Not important world story
○ Wrong category (not world)
1) select category
Comments
Save
Save and Next
2) save judgment

Category: world

Date: 2008-1-22 (Tuesday)

"WRAPUP 1-India bird flu could get out of hand -official"

Content

By Bappa Majumdar KOLKATA, India, Jan 22 (Reuters) - An outbreak of bird flu in India's most densely populated state could spiral out of control, officials said on Tuesday, as the disease spread to a seventh district. The deadly H5N1 strain of bird flu was found among poultry in Malda, infecting seven of the 19 districts in the eastern state of West Bengal, according to state officials. 'There is every chance of the virus spiraling out of hand if it's too late,' said Sanchita Bakshi, the state health services director. Nepal meanwhile banned the import of poultry products from neighbour India. In Bangladesh, authorities culled thousands of chickens after the virus spread to Natore district. At least 24 million people live in West Bengal's seven affected districts. Officials worry the virus could spread to humans and were collecting random blood samples from villagers. Experts say the H5N1 strain could mutate into a form easily transmitted from person to person, leading to a pandemic. 'We are taking all precautionary measures and

Figure 1: A screenshot of the external judging interface shown to workers within the instructions.

line and article content of the news story. According to best practices in crowdsourcing, we had three individual workers perform each HIT [12]. From these three judgments we take the majority vote for each story to create the label. The entire task totals 24,000 story judgments spread over 750 HIT instances. We paid our workers $0.50 (US dollars) per HIT (32 judgments), totalling $412.50 (including Amazon's 10% fees).

Notably, each HIT requires 32 judgments to be made, much larger than typical MTurk HITs. The reasoning behind this decision is two-fold. Firstly, the relative nature of importance in this context requires that the worker hold some background knowledge of the other news stories of the day when judging. To this end, we asked that workers make two passes over the stories. During the first and longer pass, the worker would judge each story based on the headline and content of that story and the previous stories judged, while upon the second pass, the worker can change their judgement for any story now that they have knowledge of more news stories from that day. The second reason is one of best practice. In particular, when submitting large jobs with thousands of required judgments, it has been shown that it is advantageous to retain workers over many judgments to maintain consistency in judging [11]. By increasing the HIT size, we have each worker perform at least 32 judgments.

2.2 Judging Interface

Another notable aspect of our crowdsourcing strategy was the use of an externally hosted interface. Figure 1 shows an instance of the external interface for a single HIT. Again, the reasoning was two-fold. Firstly, our previous experiences with crowdsourcing indicates that there were bots exploiting common HIT components, e.g. single entry radio buttons/text boxes, to attempt jobs on MTurk [11]. The degree of user interaction that our external interface requires makes this unlikely to be an issue. Secondly, this interface was central to our validation strategy for the work produced. Indeed, instead of using a typical validation based upon a gold-standard judgments [12], we used colour-coded summaries of the stories and the judgments that each worker made to manually validate whether they were doing an acceptable job. In particular, we qualitatively assessed each of the 750 HIT instances based on 3 criteria, namely: *1)* are all 32 stories judged, *2)* are the judgments similar across the 3 redundant judgments and *3)* are the stories marked important sensible. Although this validation strategy appears to involve a considerable volume of work, we estimate that it took no longer than 5 hours for one person to validate all 750 HIT instances, which is comparable to the time required to create a recommended gold-standard set of 5% of the full workload size. This speed is due to the fact that colour coding of the judgments factilitate assessment of criteria 1) and 2) at 'a glance', while only a small proportion of judgments need be examined under 3). Moreover, this approach is advantagious, both because one does not have to waste judgments on validation, and by manualy assessing we can have greater confidence that the workers are judging correctly. Indeed, overall the assessed work was of good quality, with less than 5% of HITs rejected.

Lastly, following an iterative design methodology [3], we submitted our HITs in 6 distinct batches, allowing for feedback to be accumulated and HIT improvements to be made. Indeed, between each batch we made minor modifications to the judging interface and updated the instructions based upon feedback from the workers.

3. EVALUATING CROWDSOURCED RELEVANCE JUDGMENTS

In this section, we analyse our crowdsourcing job and the relevance assessments produced. We aim to determine how successful crowdsourcing was and areas where improvements can be made. In particular, in each of the following four subsections, we investigate a research question. These are:

1. Is crowdsourcing actually fast and cheap? (Section 3.1)

2. Are the resulting relevance assessments of sufficient quality for crowdsourcing to be an alternative to traditional TREC assessments? (Section 3.2)

3. Is having three redundant workers judge each story necessary? (Section 3.3)

4. If we use worker agreement to introduce multiple levels of story importance, would this affect the final ranking of systems at TREC? (Section 3.4)

3.1 Crowdsourcing Analysis

Before analysing the actual relevance assessments produced, it is useful to examine the salient features of the crowdsourcing job. In particular, it has been suggested that the crowdsourcing of relevance assessments can be completed at little cost, and often very quickly [3]. We investigate whether this was indeed the case for our TREC task.

Batch	# Query Days	# HITs	# judgments	Hourly-rate ($)
Batch 1	1	15	480	3.284
Batch 2	9	135	4320	3.574
Batch 3	10	150	4800	3.528
Batch 4	10	150	4800	5.184
Batch 5	10	150	4800	4.89
Batch 6	10	150	4800	6.056

Table 1: Average amount paid per hour to workers and work composition for each batch of HITs.

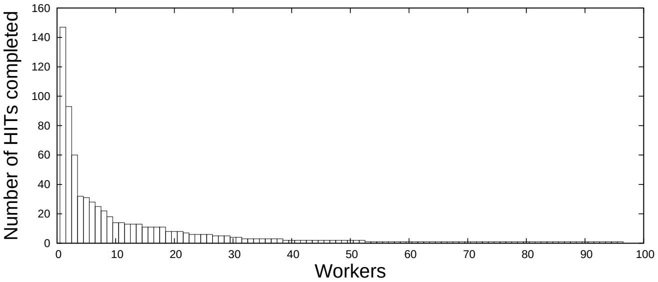


Figure 2: The number of HITs completed by each of our workers.

Prior to launching our job, we estimated that to judge the 32 stories (one HIT) it would take approximately 15 minutes, accounting for the one-off time to read the instructions and the time taken to read each story. Based upon an estimated hourly-rate (amount paid per hour of work completed) of $2, we paid a fixed rate of $0.50 per HIT. Table 1 reports the per-hourly rate paid to workers during each of the six batches. There are two points of interest. Firstly, our hourly-rate is higher than expected ($3.28 to $6.06), indicating that workers took less time than estimated to complete each HIT. Secondly, we observe an upward trend in the hourly-rate in later batches. This shows that in general, HITS in these batches took equal to or less time to complete (although there are exceptions). We believe that there are two reasons for this: firstly, between each batch we iteratively improved the instructions, hence making the task easier, and secondly, we observed a high degree of worker retainment between batches and, as such, the workers had the opportunity to become familiar with the task. Indeed, as can be seen from Figure 2, which reports the number of HITs completed by each of the 96 workers, the majority of the HITs were completed by only three workers.

In terms of the time taken by our batches, we observed a quick uptake by MTurk workers. For each of the 6 batches, the first HITs were often accepted within 10 minutes of launch, whilst the time to complete all HITs in each batch never exceeded 5 hours. Overall, crowdsourcing took a total of 8 working days to accumulate the 24,000 judgments required, including time taken by worker validation and interface improvements.

In general, we conclude that crowdsourcing judgments can be both inexpensive at $0.0156 per judgement and fast to complete. However, we believe that this task may be done 38% cheaper, as we paid above average rates for the work.

3.2 Relevance Assessment Quality

To determine the quality of our judgments, we measure the agreement between our workers. Table 2 reports the percentage of judgments for each relevance label and the between-worker agreement in terms of Fleiss Kappa [8], on average, as well as for each of the five news categories. In general, we observe that agreement on average is high (69%), lending confidence to the judgment quality. However, of interest is that agreement varies markedly between news categories. In particular, the Science/Technology and Sport categories exhibit the highest agreement with 83% and 78% respectively, while the U.S. and World categories show less agreement. Based upon the class distribution for these categories, the disparity in agreement indicates that distinquishing science from non-science stories is easier than for the U.S. or World categories. This is intuitive, as the U.S. and World categories suffer from a much higher story overlap. For example, for the story "President meets world leaders regarding climate change", it is unclear whether it is a World and/or U.S. story. Hence, workers may disagree whether it should recieve the 'important' or 'wrong category' label.

Majority (Official Qrels)	statMAP	1st Meta Worker	2nd Meta Worker	3rd Meta Worker
POSTECH_KLE	0.2206	ikm100	POSTECH_KLE	ICTNET
ikm100	0.2151	POSTECH_KLE	ICTNET	POSTECH_KLE
ICTNET	0.2138	ICTNET	ikm100	ikm100
UoS	0.1285	UoS	UoS	UoS
uogTr	0.1139	uogTr	uogTr	uogTr
ULugano	0.1000	ULugano	ULugano	ULugano
τ Correlation		0.8667	0.8667	0.7333

Table 3: Group rankings (based upon the best run submitted) using majority of three judgments against single judgments. The bottom row reports the Kendall's τ correlation between the majority and single worker rankings.

Overall, we conclude that based upon the high level of agreement observed, the relevance labels produced are of sufficient quality. Indeed, our agreement is greater than that observed in many studies of TREC assessments [1]. Hence, crowdsourcing appears to be a viable alternative to traditional TREC assessments for the Blog track top stories task.

3.3 Redundant judgments

In-line with best practices in crowdsourcing, we had three individual workers judge each HIT. However, it is important to determine to what extent this is necessary, as this is an area where costs can be dramatically decreased. To investigate this, we examine the effect of using only a single judgement on the ranking of groups that participated in TREC 2010. If the group ranking changes little, then quite possibly there is no need to have many workers judge each HIT. Table 3 reports the ranking of the six TREC 2010 groups (based upon their best run) when using the majority of the three workers (the official qrels) and the group ranking using the judgments produced by the three redundant workers individually. Furthermore, similarly to [13], Table 3 also reports Kendall's τ correlation between the group rankings produced by majority and single worker judgments. Interestingly, we observe that there is no change in the relative ranking of groups for the lower ranks, while there is a marked difference for the top three groups. As such, we conclude that redundant judging is necessary for this task, as the ranking of participating groups is not sufficiently stable at the top of the ranking, where the performances (shown in column 2) are closer.

3.4 Graded judgments

One of the advantages of using redundant judgments is that one can infer judgement confidence based on worker agreement. In particular, although not used during TREC 2010, we also created an alternative assessment set, where

Category	Important	Not	Wrong	Agreement
		Important	Category	(Kappa Fleiss)
U.S. News	21%	39%	40%	63.53%
World News	24%	38%	38%	51.69%
Sport	21%	29%	49%	77.67%
Business/Finance News	24%	43%	33%	66.88%
Science/Technology	4%	10%	86%	82.97%
Average	19%	31%	49%	68.55%

Table 2: Judgement distribution and agreement on a per category basis.

	statMap	statMNDCG@10	statMNDCG@10
	binary	binary	graded
statMAP binary	1.0000	0.4667	0.4667
statMNDCG@10 binary	-	1.0000	0.2000
statMNDCG@10 graded	-	-	1.0000

Table 4: Kendall's τ correlation between binary and graded relevance judgments under statMAP and statMNDCG@10 measures over the cross-category mean.

a news story's importance was measured on a three level graded scale [13]. In particular, if all workers judged a story important then the story was assigned a new 'highly important' label, two out of three workers resulted in an 'important' label, while one or no workers resulted in a 'not important' label, again following worker majority. This differs from the official binary qrels that distinguish 'important' from 'not important' only. In this section, we examine how the two level (binary) judgments compare to this three-level graded alternative. We aim to determine whether using this additional agreement evidence adversely affects the ranking of the TREC 2010 participants.

Table 4 reports Kendall's τ correlation between the ranking of groups under the binary and graded relevance judgments using the statMAP and statMNDCG@10 evaluation measures [4]. A high correlation indicates that the participating groups were not affected by the addition of a 'highly relevant' category, while a low correlation indicates that some groups favoured highly relevant stories more than others. From Table 4 we observe that the rankings produced by the binary and graded relevance assessments are not particularly well correlated, especially under statMNDCG@10. This indicates that the group ranking is affected by the addition of a highly relevant category. We believe that this merits further investigation, which we leave for future work.

4. CONCLUSIONS AND BEST PRACTICES

In this paper, we have described our crowdsourcing approach for creating relevance judgments for the TREC 2010 Blog track top news stories identification task. Based upon the high levels of agreement between our workers in addition to the manual validation that we performed, we believe that crowdsourcing is a highly viable alternative to TREC judging. Furthermore, we have confirmed the importance of redundant judging for relevance assessment in a TREC setting and shown that expanding the binary relevance assessments using worker agreement can strongly affect the overall ranking of participating groups. Indeed, we believe that this is an interesting area for future work.

Based upon our successful experience in crowdsourcing, we recommend the following four best practices in addition to those documented in [11], both for organisers of future TREC tracks considering a crowdsourced alternative, but also for the wider crowdsourcing community:

1. **Don't be afraid to use larger HITs:** As long as the workers perceive that the reward is worth the work, uptake on the jobs will still be high.

2. **If you have an existing interface, integrate it with MTurk:** There is often no need to build a new evaluation for MTurk, with a few tweaks and sufficient instruction, workers can use existing software.

3. **Gold-judgments are not mandatory:** While worker validation is essential, there are viable alternatives. We successfully validated all HITs manually with the aid of colour-coded summaries.

4. **Re-cost your HITs as necessary:** As workers become familiar with the task they will become more proficient and will take less time. You may wish to revise the cost of your HITs accordingly if cost is an issue.

5. REFERENCES

[1] A. Al-Maskari, M. Sanderson and P. Clough Relevance judgments between TREC and Non-TREC assessors. In *Proceedings of SIGIR'08.*

[2] O. Alonso and S. Mizzaro. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *Proceedings of The Future of IR Evaluation*, 2009.

[3] O. Alonso, D. E. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.

[4] J. A. Aslam and V. Pavlu. A practical sampling strategy for efficient retrieval evaluation. Technical report, North Eastern University, 2007.

[5] J. Atwood. Is Amazon's Mechanical Turk a failure?, 2010. `http://www.codinghorror.com/blog/2007/04/is-amazons-mechanical-turk-a-failure.html`, accessed on 02/06/2010.

[6] C. Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of EMNLP'09.*

[7] J. Downs, M. Holbrook, S. Sheng, and L. Cranor. Are your participants gaming the system? Screening Mechanical Turk workers. In *Proceedings of CHI'10.*

[8] J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[9] J. Howe. The rise of Crowdsourcing, 2010. `http://www.wired.com/wired/archive/14.06/crowds.html`, accessed on 02/06/2010.

[10] I. Ounis, C. Macdonald, and I. Soboroff. Overview of the TREC 2010 Blog Track. In *Proceedings of TREC'10.*

[11] R. McCreadie, C. Macdonald, and I. Ounis. Crowdsourcing a news query classification dataset. In *Proceedings of CSE'10.*

[12] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP'08.*

[13] E. M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of SIGIR'01.*