MSRC at TREC 2011 Crowdsourcing Track

Paul Bennett[1] **Ece Kamar**[1] **Gabriella Kazai**[2]

[1] Microsoft Research, Redmond, USA
[2] Microsoft Research, Cambridge, UK

Abstract

Crowdsourcing useful data, such as reliable relevance labels for pairs of topics and documents, requires a multidisciplinary approach that spans aspects such as user interface and interaction design, incentives, crowd engagement and management, and spam detection and filtering. Research has shown that the design of a crowdsourcing task can significantly impact the quality of the obtained data, where the geographic location of crowd workers was found to be a main indicator of quality. Following this, for the Assessment task of the TREC crowdsourcing track, we designed HITs to minimize attracting spam workers, and restricted participation to workers in the US. As an incentive, we included the possibility of a bonus pay of $5 for the best performing workers. When crowdsourcing relevance judgments, multiple judgments are typically obtained to provide greater certainty as to the true label. However, combining these judgments by a simple majority vote not only has the flawed underlying assumption that each assessor has comparable accuracy but also ignores the impact of topic specific effects (e.g. the amount of topic-expertise needed to accurately judge). We provide a simple probabilistic framework for predicting true relevance from crowdsourced judgments and explore variations that condition on worker and topic. In particular, we focus on the topic conditional model that was our primary submission for the Consensus task of the track.

1 Introduction

Crowdsourcing [7] as an online practice is increasingly adopted across a broad range of application areas, from advertising and data gathering to crisis response, design innovation, and problem solving. The term *crowdsourcing*, as coined by Jeff Howe, describes the act of outsourcing work to a large group of people or a community–a *crowd* [7]. It is an open call for contributions from members of the crowd in order to solve a problem or complete a task, often in exchange for micro-payments, social recognition, or entertainment.

A specific area where crowdsourcing can provide the required scale and efficiency is the comparative evaluation of search engines [1–3, 6, 10]. Indeed, the crowdsourcing of relevance judgments is receiving growing attention as a possible solution to enable search engine evaluation at a very large scale. However, crowdsourcing, and more specifically crowdsourcing when monetary incentives are involved, is a solution with its own set of challenges [12, 15]. Indeed, crowdsourcing has been widely criticized for its mixed quality output. Marsden, for example, argues that 90% of crowdsourcing contributions are rubbish [14]. On the other hand, several studies in relevance data collection concluded that crowdsourcing leads to reliable labels [2, 6]. At the same time, works such as [12, 23] provide evidence of cheating and random behavior among members of the crowd. Clearly, the gathering of useful data requires not only technical capabilities, but also sound experimental design. This is especially important in crowdsourcing where the interplay of the various motivations and incentives affects the quality of the collected data [15, 16].

Prompted by this growing interest in crowdsourcing for search evaluation, the new crowdsourcing track was launched at TREC with the aim to study crowdsourcing approaches for search engine evaluation. Two tasks were defined to be investigated in the first year of the track:

- ▷ The Assessment task investigates the effectiveness of crowdsourcing methods to gather relevance labels. In this task, participating teams were asked to collect topical relevance judgments from crowd workers for a small set of topics over a subset of the ClueWeb09 collection using any crowdsourcing approach, design, incentives, and platforms.
- ▷ In the Consensus task, teams were asked to compute consensus over a fixed data set containing 89k previously crowdsourced labels for a set of 19k topic-document pairs.

We participated in both the tasks. Our goal for the first task was to minimize attracting spam workers through restricting workers by geographic location and by employing more interactive user interface controls, e.g., replacing standard radio buttons, which may attract more random clicking behavior [5], with drag and drop interaction models. For the second task, we provide a probabilistic framework for predicting true relevance labels from crowdsourced judgments and explore variations that condition on worker and topic.

2 Related Work

Over the last few years, crowdsourcing has attracted a lot of attention as a valuable approach to harness human abilities from a large population of workers. A significant portion of crowdsourcing efforts has focused on consensus tasks for which a crowdsourcing system collects multiple noisy reports from workers to identify a truth about the world. Examples of consensus tasks can be found in games with a purpose (e.g., image labelling in the ESP game) [20], in citizen science projects (e.g., galaxy labelling in Galaxy Zoo) [13], and in paid crowdsourcing systems (e.g., relevance judgment for topic-document pairs) [1]. One of the primary challenges in solving consensus tasks with crowdsourcing is the recovery of the true relevance signal from the noise, i.e., both the unintentional errors and the malicious behavior in the way workers report for a consensus task [11].

There has been previous work on empirically evaluating the accuracy of non-expert workers when they report to consensus tasks and the factors that may affect their accuracy [8, 15, 17]. In particular, Alonso and Baeza-Yates, and Kazai present an empirical evaluation of the accuracy of non-expert workers in providing relevance judgments for document-topic pairs [1, 9]. In a related line of work, researchers explored approaches for learning worker models [4, 22]. Finally, there has been previous work on predicting consensus based on multiple noisy reports of workers. Previous approaches to this problem include majority voting, naive Bayes classifiers, and unsupervised and semi-supervised learning techniques [18, 19, 22].

3 Assessment Task

This section deals with the problem investigated in the Assessment task of the track, that is the effective gathering of relevance labels for a fixed set of topic-document pairs.

3.1 Task Description

The task was to collect topical relevance labels for 2175 documents over 25 topics as effectively as possible. Effectiveness was defined in terms of the quality of the workers attracted to the task, measured primarily based on the quality of the labels contributed by the individual workers. Since all crowdsourced labels had to be submitted, regardless whether a given label was later identified as being of poor quality, it was important to try to minimize attracting so-called spammers to the task.

Our team was assigned a total of 2175 topic-document pairs to judge, 520 pairs in the "assigned" set, which was shared with another 4 teams (different sets of teams per topic) and the same 1655 topic-document pairs that was shared across all participating teams. All the data was provided in sets of 5 documents per given topic, where the track guidelines required that a given worker judge all 5 pairs in a set. The data also contained gold set labels for 395 pairs, which were all grouped into sets of 5 as well, giving a total of 58 sets. Since the gold set was not distributed among the other topic-document pairs, we decided to randomly redistribute them among the total of 435 sets, thus ending up with 6 topic-document pairs per set.

3.2 Approach

Our goal was to evaluate and compare an interactive HIT design with a traditional design based on multiple radio buttons to obtain relevance labels from crowd workers. This was motivated by previous research that found significant differences in the quality of crowdsourced labels between two HIT designs, differing in the richness of the employed quality controls [10]. Other research has found that more simple designs and more mundane tasks are more susceptible to spam [5].

Our hypothesis was that more interactive HIT designs would lead to reduced spam behavior, which was shown to be more prevalent in designs that rely on multiple choice radio buttons [5]. Our interactive design is shown in Figure 1 and our baseline design is shown in Figure 2. In the interactive design, workers were asked to drag and drop the thumbnails of the web pages to be judged onto a two dimensional grid, where the horizontal placement indicates the usefulness of the page to the topic, in the worker's opinion, while the vertical placement reflects how certain the worker is in their rating of the web page. In the baseline design, we rely on a series of multiple choice inputs. These reflect the two dimensions of usefulness and confidence as in the interactive design. In addition, in the baseline, we also asked workers to pick the best document out of the 6 shown in a HIT. In both designs, workers had to click on the thumbnails to see the web pages as rendered images, as provided by the track organizers.

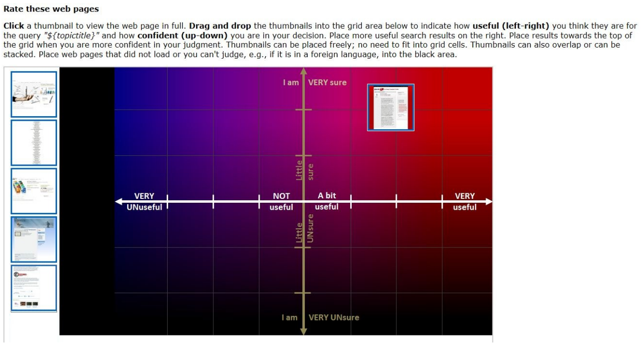
We paid $0.15 per HIT and offered a $5 bonus to the best performing workers in the event that we win the challenge. Given the 435 sets of data that needed to be judged, where one set was allocated into one HIT, and that we asked 3 workers to judge each HIT, our total cost per experiment (design) was $195.75 (without the bonus payments).

Following on from the findings in [10], we restricted participation to workers located in the US, and with a HIT approval rate of over 85%, and with a minimum of 50 completed HITs.

For each topic, we showed the title, description and narrative fields.

In addition to the relevance judgments, we also collected self-reported information on the worker's knowledge of the topic being judged, on their Big Five personality traits, whether they enjoyed the task, and if they wanted to be considered for the bonus of $5 in case we win the challenge.

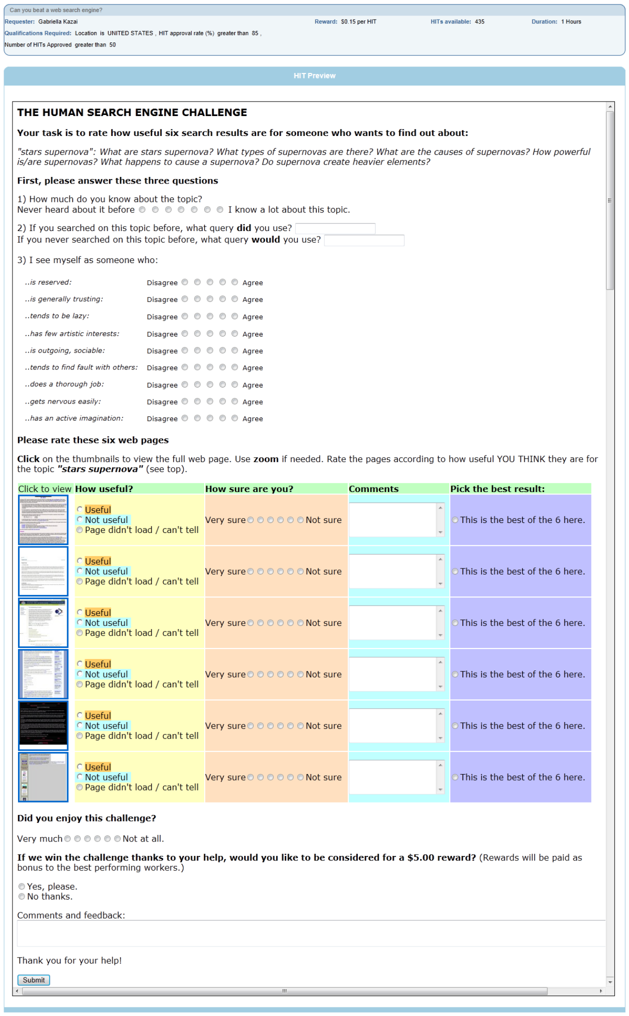
Unfortunately, due to issues with cross site scripting, where the communication between our iframe and Amazon's



Rate these web pages

Click a thumbnail to view the web page in full. **Drag and drop** the thumbnails into the grid area below to indicate how **useful (left-right)** you think they are for the query *"${topictitle}"* and how **confident (up-down)** you are in your decision. Place more useful search results on the right. Place results towards the top of the grid when you are more confident in your judgment. Thumbnails can be placed freely; no need to fit into grid cells. Thumbnails can also overlap or can be stacked. Place web pages that did not load or you can't judge, e.g., if it is in a foreign language, into the black area.

Figure 1: Drag and drop based HIT template design (showing only the drag and drop part)



Can you beat a web search engine?

Requester: Gabriella Kazai | Reward: $0.15 per HIT | HITs available: 435 | Duration: 1 Hours

Qualifications Required: Location is UNITED STATES , HIT approval rate (%) greater than 85 , Number of HITs Approved greater than 50

HIT Preview

THE HUMAN SEARCH ENGINE CHALLENGE

Your task is to rate how useful six search results are for someone who wants to find out about:

"stars supernova": What are stars supernova? What types of supernovas are there? What are the causes of supernovas? How powerful is/are supernovas? What happens to cause a supernova? Do supernova create heavier elements?

First, please answer these three questions

1) How much do you know about the topic?
Never heard about it before ○ ○ ○ ○ ○ ○ ○ I know a lot about this topic.

2) If you searched on this topic before, what query **did** you use?
If you never searched on this topic before, what query **would** you use?

3) I see myself as someone who:

..is reserved:	Disagree	○	○	○	○	○	Agree
..is generally trusting:	Disagree	○	○	○	○	○	Agree
..tends to be lazy:	Disagree	○	○	○	○	○	Agree
..has few artistic interests:	Disagree	○	○	○	○	○	Agree
..is outgoing, sociable:	Disagree	○	○	○	○	○	Agree
..tends to find fault with others:	Disagree	○	○	○	○	○	Agree
..does a thorough job:	Disagree	○	○	○	○	○	Agree
..gets nervous easily:	Disagree	○	○	○	○	○	Agree
..has an active imagination:	Disagree	○	○	○	○	○	Agree

Please rate these six web pages

Click on the thumbnails to view the full web page. Use **zoom** if needed. Rate the pages according to how useful YOU THINK they are for the topic ***"stars supernova"*** (see top).

Click to view	How useful?	How sure are you?	Comments	Pick the best result:
	○ Useful ○ Not useful ○ Page didn't load / can't tell	Very sure ○ ○ ○ ○ ○ ○ Not sure		○ This is the best of the 6 here.
	○ Useful ○ Not useful ○ Page didn't load / can't tell	Very sure ○ ○ ○ ○ ○ ○ Not sure		○ This is the best of the 6 here.
	○ Useful ○ Not useful ○ Page didn't load / can't tell	Very sure ○ ○ ○ ○ ○ ○ Not sure		○ This is the best of the 6 here.
	○ Useful ○ Not useful ○ Page didn't load / can't tell	Very sure ○ ○ ○ ○ ○ ○ Not sure		○ This is the best of the 6 here.
	○ Useful ○ Not useful ○ Page didn't load / can't tell	Very sure ○ ○ ○ ○ ○ ○ Not sure		○ This is the best of the 6 here.
	○ Useful ○ Not useful ○ Page didn't load / can't tell	Very sure ○ ○ ○ ○ ○ ○ Not sure		○ This is the best of the 6 here.

Did you enjoy this challenge?

Very much ○ ○ ○ ○ ○ ○ Not at all.

If we win the challenge thanks to your help, would you like to be considered for a $5.00 reward? (Rewards will be paid as bonus to the best performing workers.)

○ Yes, please.
○ No thanks.

Comments and feedback:

Thank you for your help!

Submit

Figure 2: Baseline HIT template design

Mechanical Turk broke down, we failed to run the experiments with the interactive design. Thus, we only report results for our baseline run. We hope to complete the interactive runs in the near future.

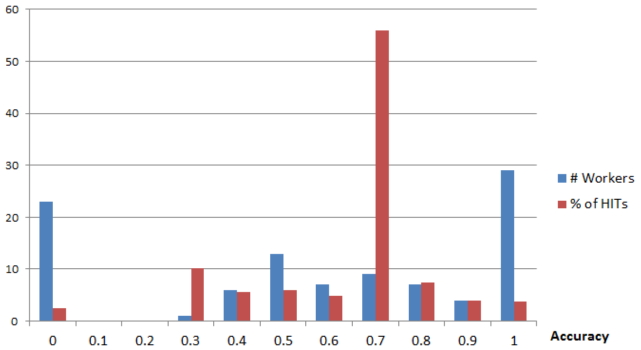


Figure 3: Number of workers and percentage of HIT volume per accuracy bin

3.3 Evaluation

A total of 1047 HITs were completed by 99 unique workers by the TREC submission deadline (the total of 1305 HITs were completed by 111 workers). As common for most crowdsourcing engagements, the majority of the HITs were completed by a few workers (62% of HITs by 4 workers). The average time workers spent on the task was 331 seconds and the average accuracy on the test questions (data with gold standard) was 0.57. Out of the four 'keen' workers, three were above average quality (accuracy levels around 0.65, with average time of 220 seconds spent per HIT), while the fourth may have been a spam worker (accuracy of 0.25, average time spent of 168 seconds). Workers reported an average familiarity level of 1.86 (the scale was 0 to 5, 0 being "never heard about the topic before" and 5 meaning "I know a lot about the topic"). Workers found the task more fun than not: average reported fun level was 3.7 (scale was 0 to 5, 0 meaning not fun at all and 5 meaning "very much enjoyed the task').'

Figure 3 shows the distribution of workers per bins of observed accuracy (on gold set). This suggests that most workers were reliable and that most of the HITs resulted in high quality data. The 23 workers with 0 accuracy only contributed a total of 26 HITs, each completing on average a single test (thus we have low certainty in the obtained accuracy scores). The only clearly 'bad apple' in the experiment was the single keen worker who completed 106 HITs and whose accuracy was 0.25.

Table 1 shows the official results against the consensus ground-truth (binary labels) and the gold set, showing the number of topic-document pairs, the number of unique workers, the Accuracy, Recall, Precision, Specificity, (normalized) Log-loss, (normalized) KL-divergence, and the root mean square error. For reference we also evaluate the consensus labels against the gold set.

We obtain a label accuracy of 77% against the consensus data, and 65% against the editorial judgments from TREC. In comparison, TREC assessors have pairwise agreement

Table 1: Assessment task results for primary runs, against gold and consensus ground-truth sets

Evaluation	Pairs	Wrks	Acc	R	P	S	LL	nLL	KL	nKL	RMSE
Consensus ground-truth	2005	95	0.77	0.71	0.86	0.83	1009.29	7.57	2551.40	17.26	0.70
Gold set ground-truth	2005	95	0.65	0.64	0.79	0.62	198.20	11.11	198.23	7.66	0.54
Consensus vs. gold set	1875	1	0.80	0.87	0.85	0.66	828.93	6.85	829.16	3.14	0.45

levels of 70–80% on average, with high variability across topics [21, p.44]. Thus, we obtain a similar agreement level to that reported in TREC for the evaluation against the consensus labels, but agreement with the gold set labels is lower. The difference in the two different evaluations is relatively large, but we can also see that the consensus labels obtain only a label accuracy of 80% when evaluated against the gold set.This can raise questions regarding the use of consensus or gold standard as ground-truth.

The precision, 86% against consensus and 79% against the gold set, is higher than the recall, 71% against consensus and 64% against the gold set, indicating that the design led to judgments based on a relatively strict relevance criterion. Plausible design factors impacting this are the following. First, we showed the full narrative of the topic, restricting the topic to the interpretation of the original topic creator. Second, we used an operational definition of "topical relevance", defining it in terms of "usefulness", which may be regarded as a more restrictive criterion—not all topically relevant documents are useful. Third, we also asked workers to indicate their confidence in their ratings, which may have averted more speculative guessing.

4 Consensus Task

This section deals with the problem that is investigated in the second part of the TREC challenge of predicting consensus of a set of noisy worker judgments collected for a relevance judgment task. We start by introducing the consensus task and the dataset that is used for studying the consensus task. We then present a short survey of related work on solving consensus tasks in the domain of query judgments and in other domains. Next, we propose multiple approaches for solving consensus tasks that fall out of the same probabilistic formulation by making different independence assumptions. Finally, we conclude the section by evaluating these approaches on the relevance judgment dataset provided by the organizers.

4.1 Task Description

The second part of the TREC Crowdsourcing Track, referred to as the "consensus task", focuses on computing consensus about the relevance of a given document to a topic over a set of individual reports collected from multiple workers. The consensus task aims at recovering the actual relevance of a topic-document pair based on multiple worker reports, thus eliminating the noise in the way workers judge relevance. Supervised, semi-supervised and unsupervised learning approaches can be used for predicting consensus from a set of noisy worker reports. In this paper, we propose using supervised learning approaches based on the naive Bayes method to predict relevance.

We use a dataset provided by the organizers of the challenge to train predictive models and to evaluate the prediction accuracy of these models. The dataset is composed of individual worker judgments of topic-document pairs, which were collected as part of the TREC 2010 Relevance Feedback track [19]. A small subset of instances in the dataset has ground truth NIST judgments. Each topic-document pair is judged by a worker as a part of a human intelligence task (HIT) advertised on the Mechanical Turk platform. Each instance of the dataset is composed of identifiers for a given topic and document pair, an identifier for the HIT in which the relevance is collected, an identifier for the worker who judged the pair, the judgment of the worker (relevant/not relevant) and the ground truth judgment of the pair. The dataset includes 89,624 relevance judgments collected from 762 workers for 19,033 topic-document pairs. A subset of this dataset was released as the development (training) set. The training set includes a total of 10,770 judgments collected from 181 unique workers. Approximately 15% of the training dataset has ground truth NIST judgments. The majority of the document-topic pairs have 3 judgments, and the remaining small subset of the training set has 6 judgments.

The track evaluation uses both the ground truth NIST judgments (gold set) for evaluation as well as the consensus label computed from other participating teams' predictions for topic-document pairs in the test set. In this report, we focus on results over the gold set in the development set and over the gold set within the test set. Section 4.3 presents a set of results from these empirical evaluations.

4.2 Approach

The dataset provided for the TREC challenge includes labeled and unlabeled relevance judgment instances; approximately 15% of the instances include gold standard relevance judgments. In a crowdsourcing environment, often some of the examples have been labeled by experts to either serve as teaching items for workers or to act as honeypots, challenges, etc. We define the problem of predicting whether a document is relevant to a topic as a supervised learning problem. We train predictive models with the subset of the data instances that have expert or "gold" labels. We first define the learning problem of predicting relevance. Then, we present different approaches for predicting the relevance of a document to a topic.

We build models based on various assumptions. These models correspond to three broad conditions: (1) an assumption that workers have comparable accuracy across all tasks; (2) an assumption that workers have comparable accuracy withing a task, but varying across tasks (e.g. the amount of topic-expertise needed to accurately judge is the primary determiner of worker accuracy); (3) each worker has a particular skill/accuracy in making relevance judgments across all tasks.

Ultimately, we desire to infer the probability of the actual relevance label of a document, taking into account topic-specific effects, document-specific effects, and worker-specific effects. That is, we wish to compute:

$$\Pr(R_{i,j} \mid t_i, d_j, \{w_1, \dots, w_n \mid w_k \text{is elicited for } i, j\}) \quad (1)$$

Here $t_i \in T$ is a particular topic, $d_j \in D$ is a particular document, $R_{i,j} \in \{0, 1\}$ is the event that document d_j is relevant to topic t_i, and $w_k \in \{0, 1\}$ is the relevance label provided by worker k (out of n total workers across all topics). We will abbreviate the worker labels elicited for a pair as $\vec{w}_{i:j} = w_{1,i:j}, \dots, w_{n,i:j}$ to simplify. Depending on what independence assumptions we make when computing this term, the majority of topic and worker specific effects can be captured.[1]

4.2.1 Naive Bayes Approach

The first approach we take to predict relevance is the naive Bayes approach. This approach applies Bayes' theorem and makes strong independence assumptions between the features that are used to predict relevance. In particular, it assumes that a given document, a given topic and judgements obtained from workers for the document-topic pair are independent given the relevance of the document to the topic. The way $\Pr(r_{i,j} \mid t_i, d_j, \vec{w}_{i:j})$ is computed with the naive Bayes approach is given below:

$$\Pr(r_{i,j} \mid t_i, d_j, \vec{w}_{i:j})$$

By Bayes rule

$$\propto \Pr(\vec{w}_{i:j} \mid r_{i,j}, t_i, d_j) \Pr(r_{i,j} \mid t_i, d_j)$$

Assuming the relevance prior is independent of topic and document

$$= \Pr(\vec{w}_{i:j} \mid r_{i,j}, t_i, d_j) \Pr(r_{i,j})$$

Assuming conditional independence given relevance

$$= \Pr(r_{i,j}) \prod_{k=1}^{|\vec{w}_{i:j}|} \Pr(w_{k,i:j} \mid r_{i,j})$$

Assuming exchangeability among judges

$$= \Pr(r_{i,j}) \prod_{w_k \in |\vec{w}_{i:j}} \Pr(w_k \mid r_{i,j}) \quad (2)$$

Here $\Pr(r_{i,j})$ is simply the probability of seeing a relevant document in the training set. $\Pr(w_k \mid r_{i,j})$ is the probability of a worker will say relevant/irrelevant conditional on the actual document's relevance. These probabilities are computed from the subset of the training set with gold standard relevance judgements.

4.2.2 Topic-Conditional Naive Bayes

Next, we present an approach for relaxing the independence assumptions of the naive Bayes model. The topic-conditional naive Bayes model assumes that a given document and judgements obtained from workers are independent given a topic and the relevance of the document to the topic. This model takes into account that the prior probability of a document being relevant to a topic depends on the topic, and also takes into account that the likelihood of a worker reporting relevant or not may depend on the topic. The way $\Pr(r_{i,j} \mid t_i, d_j, \vec{w}_{i:j})$ is computed with the topic-conditional naive Bayes approach is given below:

$$\Pr(r_{i,j} \mid t_i, d_j, \vec{w}_{i:j})$$

By Bayes rule

$$\propto \Pr(\vec{w}_{i:j} \mid r_{i,j}, t_i, d_j) \Pr(r_{i,j} \mid t_i, d_j)$$

Assuming the relevance prior primarily depends on topic

$$= \Pr(\vec{w}_{i:j} \mid r_{i,j}, t_i, d_j) \Pr(r_{i,j} \mid t_i)$$

Assuming conditional independence given topic and relevance

$$= \Pr(r_{i,j} \mid t_i) \prod_{k=1}^{|\vec{w}_{i:j}|} \Pr(w_{k,i:j} \mid r_{i,j}, t_i)$$

Assuming exchangeability among judges

$$= \Pr(r_{i,j} \mid t_i) \prod_{w_k \in |\vec{w}_{i:j}} \Pr(w_k \mid r_{i,j}, t_i) \quad (3)$$

Here $\Pr(r_{i,j} \mid t_i)$ is simply the probability of seeing a relevant document for this topic. While $\Pr(w_k \mid r_{i,j}, t_i)$ is the probability within this topic that a worker will say relevant/irrelevant conditional on the actual document's relevance. Both of these can be computed from the known gold set for a topic.

4.2.3 Worker-Conditional Naive Bayes

A second approach to relaxing the independence assumptions of the naive Bayes model is reasoning about the fact that the way workers report may differ from worker to worker. The worker-conditional naive Bayes model assumes

[1] If elicitation is not random, then a vector of elicitation variables should be explicitly added.

that a given document, a given topic and the relevance reports obtained from workers are independent given the relevance of the document to the topic and the history of workers in the training set. While calculating the likelihood of a worker reporting relevant for a relevant (or irrelevant) document-topic pair, this model computes the ratio of instances in which the same worker reported relevant to a relevant (or irrelevant) document-topic pair in the training set. To compute relevance probabilities with this model, we introduce a new feature h_k. h_k represents the reporting history in the training set of the worker reporting $w_{k,i:j}$ for the current task. $\vec{h}$ includes the histories of all workers reporting for the current task. The way $\Pr(r_{i,j} \mid t_i, d_j, \vec{w}_{i:j})$ is computed with the worker-conditional naive Bayes approach is given below:

$$
\begin{aligned}
& \Pr(r_{i,j} \mid t_i, d_j, \vec{w}_{i:j}, \vec{h}) \\
& \text{By Bayes rule} \\
& \quad \propto \quad \Pr(\vec{w}_{i:j} \mid r_{i,j}, t_i, d_j, \vec{h}) \Pr(r_{i,j} \mid t_i, d_j, \vec{h}) \\
& \text{Assuming the relevance prior is independent of topic,} \\
& \text{ document and history} \\
& \quad = \quad \Pr(\vec{w}_{i:j} \mid r_{i,j}, t_i, d_j, \vec{h}) \Pr(r_{i,j}) \\
& \text{ Assuming conditional independence given worker} \\
& \text{history and relevance} \\
& \quad = \quad \Pr(r_{i,j}) \prod_{k=1}^{|\vec{w}_{i:j}|} \Pr(w_{k,i:j} \mid r_{i,j}, h_k) \\
& \text{Assuming exchangeability among judges with the} \\
& \text{same history} \\
& \quad = \quad \Pr(r_{i,j}) \prod_{w_k \in |\vec{w}_{i:j}} \Pr(w_k \mid r_{i,j}, h_k) \qquad (4)
\end{aligned}
$$

Here, $Pr(w_k \mid r_{i,j}, h_k)$ is the probability that a worker will say relevant/irrelevant conditional on the actual document's relevance and the history of this worker in the training set. This probability is estimated from the training set by counting the number of times this worker reported relevant/irrelevant in the training set for document-topic pairs with the given relevance value. For workers that do not have judgements in the training set, we used a general worker history which includes judgements from all workers in the training set.

4.3 Evaluation

During the development phase, we randomly split the development data into a train (80%) and a validation portion (20%) by topic-document ID. That is, all of the ratings for a topic-document ID were either completely in the training set or completely in the testing set. Table 2 presents results computed by estimating model parameters over the training split of the development data and estimating performance on the validation portion of the development data. The column *DefaultAcc* presents the accuracy that can be obtained by guessing the most common class (relevant).

Table 3 presents the preliminary evaluation results provided by the track organizers over the gold portion of the test set. In the test set, 1000 documents had gold labels with 500 relevant and 500 irrelevant. The results have been sorted by the log-loss measure from the best (top of table) to the worst (bottom of table).

4.4 Discussion

From the result in Table 2, we see that only the topic-conditional naive Bayes outperformed the default model of predicting the most common class. However, all of the methods do outperform the majority vote method. The failure to outperform the default model may therefore be more of a result of the class skew in the data than an indication of the inferiority of the model. We decided to focus on accuracy as our decision criterion. This lead us to choose the topic-conditional model as the run to submit.

In Table 3, we see that the topic-conditional model (MSRC) performs in the middle of the pack in most classification measures (accuracy, precision, recall, specificity). However, in the two "soft measures" that measure the quality of probability estimates, log loss and root mean squared error (RMSE), the method is the best performer. This may indicate that the optimal threshold for the probability to make a hard classification decision may be other than the default (of 0.5). This requires further investigation to determine whether the models can be further optimized for classification procedures. Given the simplicity of the model, it is surprising that it can outperform the other submissions on the probability measures by a large margin.

4.5 Limitations

As seen in Table 2 evaluation over the development data was problematic for two reasons. First, there was a large skew in class prevalance. This made determining overall model. Second, the dataset was small – how these models perform as a function of the amount of data is an important question we intend to investigate over the test set.

5 Conclusions

For the assessment task, our results were ranked 3rd based on the accuracy metric on both the gold set and the consensus ground-truth set. This is promising and suggests that limiting workers to those with high HIT approval rate in the US is a good start. We plan to run the interactive experiments in the very near future to investigate our original research question.

Considering the simplicity of the topic-conditional naive Bayes model and its relative high performance with respect

Table 2: Results of Models Over Development Set

Model	TruePos	TrueNeg	FalsePos	FalseNeg	Accuracy	DefaultAcc	Prec	Recall	Specificity
Majority Vote	101	8	17	19	0.7517	0.8276	0.8559	0.8417	**0.3200**
naive Bayes	120	0	25	0	0.8276	0.8276	0.8276	**1.0000**	0.0000
nB Topic	115	7	18	5	**0.8414**	0.8276	**0.8647**	0.9583	0.2800
nB Worker	117	1	24	3	0.8138	0.8276	0.8298	0.9750	0.0400

Table 3: Preliminary assessment task results over gold set ground-truth – sorted by (negative) LogLoss from best (top) to worst (bottom)

Team	Accuracy	Recall	Precision	Specificity	LogLoss	RMSE
MSRC	0.64	0.70	0.62	0.58	**610.28**	**0.45**
uogTr	0.44	0.34	0.43	0.54	931.74	0.59
LingPipe	0.66	0.73	0.64	0.59	975.88	0.50
GeAnn	0.58	0.74	0.56	0.42	1150.44	0.51
UWaterlooMDS	0.67	0.78	0.64	0.57	1435.77	0.50
uc3m	**0.70**	0.75	**0.68**	**0.64**	2772.31	0.55
BUPT-WILDCAT	0.69	0.79	0.65	0.58	2901.26	0.56
TUD_DMIR	0.66	0.76	0.63	0.56	3113.10	0.58
UTaustin	0.60	**0.91**	0.56	0.30	3647.29	0.63
qirdcsuog	0.53	0.82	0.52	0.23	4338.07	0.69

to probability prediction, it offers a promising path for future development. Of particular interest is a model that accounts for both worker and topic effects simultaneously.

6 Acknowledgments

We would like to thank Emine Yilmaz and Ryen White for their valuable discussions and feedback.

References

[1] O. Alonso and R. A. Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In *Advances in Information Retrieval - 33rd European Conference on IR Research (ECIR 2011)*, volume 6611 of *LNCS*, pages 153–164. Springer, 2011.

[2] O. Alonso and S. Mizzaro. Can we get rid of TREC assessors? using Mechanical Turk for relevance assessment. In S. Geva, J. Kamps, C. Peters, T. Sakai, A. Trotman, and E. Voorhees, editors, *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 15–16, 2009.

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

[4] P. Dai et al. Artificial intelligence for artificial artificial intelligence. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[5] C. Eickhoff and A. P. de Vries. How crowdsourcable is your task? In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM 2011)*, pages 11–14. ACM, 2011.

[6] C. Grady and M. Lease. Crowdsourcing document relevance assessment with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 172–179, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[7] J. Howe. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Publishing Group, New York, NY, USA, 1 edition, 2008.

[8] P. G. Ipeirotis. Analyzing the amazon mechanical turk marketplace. *XRDS*, 17:16–21, December 2010.

[9] G. Kazai. In search of quality in crowdsourcing for search engine evaluation. In *Advances in Information Retrieval - 33rd European Conference on IR Research (ECIR 2011)*, volume 6611 of *LNCS*, pages 165–176. Springer, 2011.

[10] G. Kazai, J. Kamps, M. Koolen, and N. Milic-Frayling. Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 205–214, New York, NY, USA, 2011. ACM.

[11] A. Kittur, E. Chi, and B. Suh. Crowdsourcing user studies with Mechanical Turk. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 453–456. ACM, 2008.

[12] J. Le, A. Edmonds, V. Hester, and L. Biewald. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, pages 21–26, 2010.

[13] C. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. Raddick, R. Nichol, A. Szalay, D. Andreescu, et al. Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189, 2008.

[14] P. Marsden. Crowdsourcing. *Contagious Magazine*, 18:24–28, 2009.

[15] W. Mason and D. Watts. Financial incentives and the performance of crowds. *ACM SIGKDD Explorations Newsletter*, 11(2):100–108, 2010.

[16] A. J. Quinn and B. B. Bederson. Human computation: a survey and taxonomy of a growing field. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 1403–1412, New York, NY, USA, 2011. ACM.

[17] A. Shaw, J. Horton, and D. Chen. Designing incentives for inexpert human raters. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, CSCW '11, 2011.

[18] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263. Association for Computational Linguistics, 2008.

[19] W. Tang and M. Lease. Semi-supervised consensus labeling for crowdsourcing. In *SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval (CIR)*, 2011.

[20] L. Von Ahn and L. Dabbish. Designing games with a purpose. *Communications of the ACM*, 2008.

[21] E. M. Voorhees and D. K. Harman, editors. *TREC: Experimentation and Evaluation in Information Retrieval*. MIT Press, 2005.

[22] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in Neural Information Processing Systems*, 22:2035–2043, 2009.

[23] D. Zhu and B. Carterette. An analysis of assessor behavior in crowdsourced preference judgments. In *SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, pages 17–20, 2010.