Assessor Differences and User Preferences in Tweet Timeline Generation

Yulu Wang[1], Garrick Sherman[2], Jimmy Lin[1], and Miles Efron[2]

[1] University of Maryland, College Park
[2] University of Illinois, Urbana-Champaign

{ylwang,jimmylin}@umd.edu, {gsherma2,mefron}@illinois.edu

ABSTRACT

In information retrieval evaluation, when presented with an effectiveness difference between two systems, there are three relevant questions one might ask. First, are the differences statistically significant? Second, is the comparison stable with respect to assessor differences? Finally, is the difference actually meaningful to a user? This paper tackles the last two questions about assessor differences and user preferences in the context of the newly-introduced tweet timeline generation task in the TREC 2014 Microblog track, where the system's goal is to construct an informative summary of non-redundant tweets that addresses the user's information need. Central to the evaluation methodology is human-generated semantic clusters of tweets that contain substantively similar information. We show that the evaluation is stable with respect to assessor differences in clustering and that user preferences generally correlate with effectiveness metrics even though users are not explicitly aware of the semantic clustering being performed by the systems. Although our analyses are limited to this particular task, we believe that lessons learned could generalize to other evaluations based on establishing semantic equivalence between information units, such as nugget-based evaluations in question answering and temporal summarization.

Categories and Subject Descriptors: H.3.4 [Information Storage and Retrieval]: Systems and Software—Performance evaluation

Keywords: TREC evaluation; microblog search; user study

1. INTRODUCTION

In response to information needs, search systems should strive to return as much relevant content as possible. However, there are other factors that systems might consider beyond topical relevance: often, users may not wish to see multiple pieces of text that "say the same thing"; frequently, users desire diverse results that cover multiple aspects of the topic they are interested in. These considerations are particularly important in the context of searching tweets on Twitter. Due to the nature of the medium, there are frequently duplicate, near-duplicate, and highly-similar posts—for example, a breaking news event may be reported by multiple outlets at roughly the same time, each using slightly different language. In turn, these posts are retweeted, sometimes with additional commentary, copy-and-pasted, and discussed as part of a global conversation. This creates a cacophony of voices that obscures the underlying information content.

The Microblog track at TREC began in 2011 to explore information retrieval challenges in the context of social media services such as Twitter. The main task since 2011 has been temporally-anchored ad hoc retrieval ("At time T, give me the most relevant tweets about an information need expressed as query Q"). In response to the considerations above, the tweet timeline generation (TTG) task was introduced in 2014, where the goal is to construct a "timeline" of relevant and non-redundant tweets that best addresses the user's information need.

This paper presents a meta-evaluation of the TTG task from two perspectives: First, do assessor differences affect evaluation stability? Ultimately, TTG evaluation boils down to judgments about the semantic content of tweets, for which we would expect inter-assessor disagreements. Do these differences prevent us from drawing conclusions about the relative effectiveness of systems? Second, do user preferences correlate with our evaluation metrics? That is, are our metrics meaningful in being able to capture aspects of what users care about in timelines for making system comparisons? Our major findings are summarized as follows:

- We find that assessors do exhibit substantial differences in semantic clustering. However, these differences do not impact the stability of system comparisons.
- We find that user preferences correlate with metric differences, and for precision and unweighted F_1, agreement increases as the magnitude of the difference increases.

Although this work focuses on a specific TREC task, tweet timeline generation is representative of a class of information retrieval evaluations based on identifying atomic units of information and establishing semantic equivalences between these units. Thus, we believe that lessons learned from our study can be applied to similar types of evaluations.

2. BACKGROUND AND RELATED WORK

At a high level, the tweet timeline generation task bears a family resemblance to topic detection and tracking (TDT)

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

ACM 978-1-4503-3621-5/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767699.

[33, 3], "other" nuggets in question answering [32, 9, 8], nugget-based evaluations in DARPA's BOLT program,[1] and temporal summarization at TREC [12, 5]. These tasks all share the insight that atomic units of information should be grouped together into semantic equivalence classes to provide more useful responses to users (we refer to this generically as "clustering"). In TDT the atomic units were documents, which were grouped into those discussing the same event. For question answering, nuggets were short phrases that provided interesting information about a target entity. In temporal summarization, sentences formed the atomic units that conveyed essential information about a particular event. These notions are also reflected in variants of ad hoc retrieval such as aspect retrieval [21] (or, alternatively, facet or sub-topic retrieval [34]) and result diversification [1, 28], although less explicitly.

The two biggest challenges of formulating such an evaluation are (1) defining the atomic unit of information and (2) defining semantic equivalence between those units. Once these two issues are resolved, computing a metric is reasonably straightforward, although variations abound. The first issue is more difficult than it initially seems if the atomic unit does not correspond to a logical entity such as a document. Experience from question answering evaluations has shown that users disagree about the granularity of nuggets—for example, whether a piece of text encodes one or more nuggets and how to treat partial semantic overlap between two pieces of text [20]. The solution that most evaluations adopt today is to simply declare the atomic information unit *by fiat*: in tweet timeline generation, tweets form the atomic units, and in temporal summarization [5], they are sentences. This sweeps many nuances under the rug, but yields workable evaluations in practice.

The second issue—semantic equivalence between atomic information units—is challenging because making such judgments requires taking into account context and fine-grained distinctions in meaning. In the same way that assessors disagree over relevance judgments (see [6] for a nice summary), humans also disagree about whether two pieces of text have the same semantic content. This issue is typically resolved by acknowledging these assessor differences and simply accepting the opinion of a *single* assessor. In the same way that Voorhees [31] demonstrated the stability of system rankings with respect to assessors' divergent document-level relevance judgments, the implicit assumption is that for these semantic clustering tasks, assessor differences don't matter if our goal is to induce a set of system comparisons.

To our knowledge, however, this assumption has never been validated at scale. For example, "ground truth" nuggets for question answering evaluations were generated by a single assessor (for each entity) [32, 9].[2] Thus, it is unclear if an alternative set of nuggets (i.e., generated by another assessor) would have altered system rankings. A more recent iteration of the nugget-based evaluation methodology [22] has not to our knowledge examined assessor differences in the "nuggetization" process either. One of the major contributions of this paper is that, for tweet timeline generation, we devoted the resources necessary to answer this question by procuring two independent sets of semantic clusters for all topics in the test collection. Our results verify that system rankings are stable with respect to assessor differences in the TTG task, which gives us some confidence that the same results will carry over to similar tasks as well.

With respect to the second part of our paper—an exploration of user preferences in tweet timeline generation—there is a long history of research that examines the correlation between effectiveness metrics from system-oriented evaluations with task-based metrics from user-oriented evaluations [13, 29, 4, 30, 14, 2, 26, 25, 27]. The broader goal of this thread of work is to understand when system effectiveness improvements are meaningful or useful in improving users' information seeking abilities in practice. Early studies suggest that "better" systems (as measured by system-oriented metrics) don't necessarily translate into better task performance [13, 29, 30]. Smith and Kantor [26] explained that users "do well" with poor search engines because they reformulate queries. Along these lines, Lin and Smucker [19] suggested that browsing can compensate for poor search results. However, more recent work has been able to detect a correlation between various system effectiveness metrics and human preferences [2, 25] (by using larger sample sizes) and between precision and user performance [27] (by more carefully controlling the experimental setup). These results give the community some degree of confidence that system-oriented evaluations can guide progress in producing more useful systems.

Our study follows the general setup of Sanderson et al. [25], i.e., asking users which of two system outputs they prefer. There are, however, a few important differences. Whereas their work examined ad hoc retrieval, we explore the more complex task of tweet timeline generation. Whereas they restricted system comparisons to those with "large" differences and "small" differences in effectiveness, we sampled our comparison conditions and analyzed the results in a way that allows us to characterize user sensitivity to different magnitudes of differences. Finally, whereas Sanderson et al. used judgments from Amazon's Mechanical Turk service, we took the route of training local assessors,

3. TWEET TIMELINE GENERATION

3.1 Task Definition

Tweet timeline generation (TTG) was introduced at the TREC 2014 Microblog track to supplement the existing ad hoc retrieval task. The putative user model is as follows: "At time T, I have an information need expressed by query Q, and I would like a summary that captures relevant information." The system's task is to produce a "summary" timeline, operationalized as a list of non-redundant, chronologically ordered tweets. It is imagined that the user would consume the entire summary (unlike a ranked list, where the user might stop reading at any time).

We conceived of a reference architecture in which a TTG module processes the output of an ad hoc retrieval system to generate the timeline. Thus, tweet timeline generation introduces two additional challenges beyond ad hoc retrieval:

- Systems must detect (and eliminate) redundant tweets. This is equivalent to saying that systems must detect if a tweet contains novel information.

[1] `http://www.nist.gov/itl/iad/mig/bolt_p1.cfm`

[2] Subsequent "nugget pyramid" extensions [17, 9, 8] had multiple assessors assigning importance *weights*, but the underlying nuggets were still the same; similarly, researchers have studied how assessors identify nuggets in a piece of text [20], but the analyses were restricted to a *single* set of nuggets.

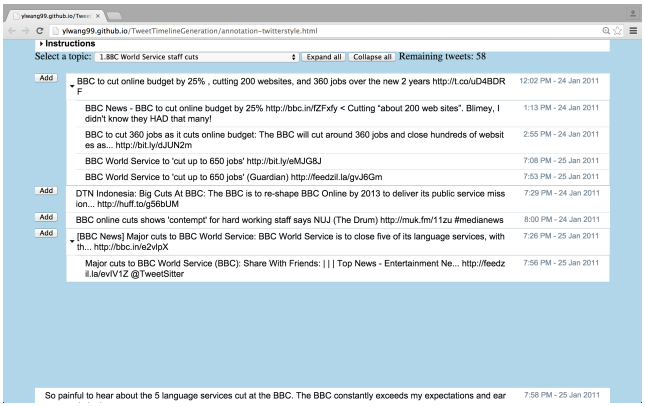

Figure 1: Screenshot of the annotation interface. Tweets are presented one at a time in chronological order (bottom). For each tweet, the assessor can add it to an existing cluster or create a new cluster.

- Systems must determine how many tweets to return. Some topics have more relevant and non-redundant tweets than others and a system must be able to automatically infer this. Systems can make different precision/recall tradeoffs along these lines.

We operationalized redundancy as follows: for every pair of tweets, if the chronologically later tweet contains substantive information that is not present in the earlier tweet, the later tweet is considered novel; otherwise, the later tweet is redundant with respect to the earlier one. In our definition, redundancy and novelty are antonyms, and so we use them interchangeably, but in opposite contexts.

Due to the temporal constraint, redundancy is *not* symmetric. If tweet A precedes tweet B and tweet B contains substantively similar information found in tweet A, then B is redundant with respect to A, but not the other way around. We also assume transitivity. Suppose A precedes B and B precedes C: if B is redundant with respect to A and C is redundant with respect to B, then by definition C is redundant with respect to A. In this task setup, redundancy boils down to the definition of "contains substantively similar information", which is more precisely defined below.

3.2 Annotation Methodology

The TTG definition of redundancy and the assumption of transitivity means that the task can be viewed as semantic clustering—that is, we wish to group relevant tweets into clusters in which all tweets share substantively similar information. Within each cluster, the earliest tweet is novel; all other tweets in the cluster are redundant with respect to all earlier tweets.

Our annotation methodology builds exactly on this idea. We begin with a list of all relevant tweets, ordered chronologically, from earliest to latest. These tweets are presented, one at a time, to a human assessor. For each tweet, the assessor can add it to an existing cluster if she thinks the tweet contains substantively similar information with respect to tweets in the existing cluster, or she can create a new cluster for the tweet. We have developed a JavaScript-based annotation interface to help assessors accomplish this task. A screenshot is shown in Figure 1.

In the interface, the next tweet to be clustered is shown at the bottom of the screen. The assessor can either add the tweet to an existing cluster by clicking the "Add" button next to the cluster or create a new cluster by hitting the space bar. At any time, the assessor can expand a cluster to show all tweets contained in it, or collapse the cluster to show only the first tweet. The interface also implements an undo feature that allows the assessor to reverse the action taken and go back to the previous tweet.

The TTG evaluation methodology boils down to this central question: what exactly does "substantively similar information" mean? Like document relevance in ad hoc retrieval, assessors make the final determination and we expect natural variations among humans. However, pilot studies helped us devise a set of guidelines, which were provided as instructions to the assessors. We told them: a good rule of thumb is that if two tweets "say the same thing", then they're substantively similar. To speed up the clustering process, the annotators were asked not to consider external content (e.g., follow links in the tweets).

To provide further guidance, we devised a few questions that the assessors might consider in determining whether two tweets should be in the same cluster:

- If I had already seen the first tweet, would I have missed out on some information if I didn't see the second tweet?
- If two tweets are similar but the second contains an addition to or endorsement of the first, is the addition or endorsement important enough that I would be interested in seeing both tweets?
- Sometimes two tweets look similar but actually narrate the development of an event. Are the tweets different enough from each other that I would want to see both tweets to understand how an event develops or unfolds?

Since TTG was originally conceived as a stage following ad hoc retrieval, the track guidelines asked TTG participants to also submit ad hoc runs. For the evaluation, NIST assessors developed 55 topics and used a standard pooling methodology to generate tweet-level relevance judgments; see the track overview for more details [18]. Tweets in the pool were assigned one of three judgments: not relevant, relevant, and highly relevant. In the cluster annotation process, assessors at UMD and UIUC worked on the relevant and highly-relevant tweets from the NIST judgment pools. Due to resource constraints, NIST assessors were not able to perform the clustering, and thus a weakness of our setup is that the individual with the information need was not the one who created the clusters. The two assessors at UMD were graduate students in computer science (both male). The two assessors at UIUC were graduate students in library and information science (one male, one female). All were familiar with Twitter.

Assessors were first trained in the laboratory: the session included an introduction to the task and an overview of the annotation interface. After that, assessors were free to perform annotations at their own pace on their own machines, at any location of their choosing. This was possible because the annotation interface was implemented in JavaScript and hence accessible over the web. All assessors began with a throwaway "practice topic" (although they were not aware of the throwaway nature) and then proceeded to annotate topics in batches (roughly ten topics per batch). Topics were

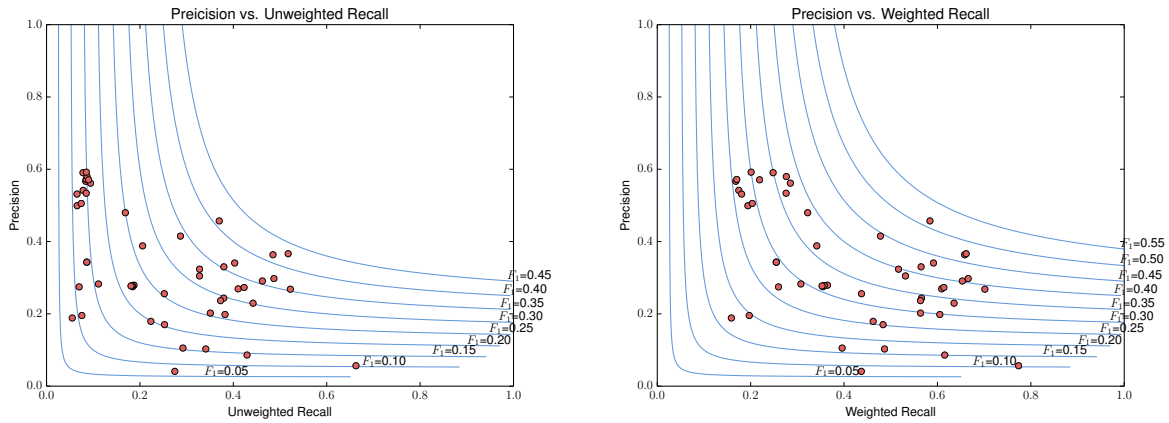

Figure 2: Scatter plots showing precision vs. unweighted recall (left) and precision vs. weighted recall (right) for all submitted runs in the TREC 2014 TTG task, overlaid with iso-F_1 contours.

grouped into batches of roughly equal size (in terms of the number of relevant tweets) prior to the beginning of the annotation process. When an assessor completed a batch, he or she could request another batch to work on.

Each site annotated the topic batches in the opposite order, and when a covering set for all topics had been obtained, we designated those clusters to be the "official" judgments. However, the annotation process continued until both sites had processed all topics, giving us alternate judgments. Thus, both the official and alternate clusters contained annotations generated from both sites.

3.3 Metrics and Results

The output of the human annotation process is an ordered list of tweet clusters. Within each cluster, the tweets are sorted by temporal order (earliest to latest). The clusters themselves are sorted by the temporal order of their earliest tweet. Following the heuristic of using the most straightforward metric when defining a new task (and then subsequently refining the metric as needed), we decided to measure cluster-based precision and recall. The measure is cluster-based in the sense that systems only receive credit for returning one tweet from each cluster—that is, once a tweet is retrieved, all other tweets in the cluster are automatically considered not relevant. From this, we can compute precision, recall, and F-score in the usual way (lacking any basis for setting the β parameter, we simply computed F_1). Since the user model assumes that the user will consume the entire summary, set-based metrics seemed appropriate.

The only additional refinement is that we computed both weighted and unweighted variants of recall. In weighted recall, each cluster is assigned a weight proportional to the sum of relevance grades from every tweet in the cluster (relevant tweets receive a weight of one and highly-relevant tweets receive a weight of two). This weighting scheme implements the heuristic that larger clusters and those containing more highly-relevant tweets are more important, and the denominator in the weighted recall computation is the sum of all cluster weights. In unweighted recall, all clusters are considered equally important, and the denominator is simply the total number of clusters.

Note that this setup gives equal credit to retrieving *any* tweet from a cluster. Intuitively, however, this seems overly simplistic—users would certainly prefer seeing certain tweets over others, even if they contain substantively similar information [15]. For example, users might prefer the earliest tweet, a tweet from the most "authoritative" user (e.g., a verified news account), or a tweet from someone close by in their network (e.g., a tweet from someone they follow). We currently do not have sufficient understanding to accurately model such preferences, and thus explicitly made the decision not to tackle this challenge.

The evaluation metrics for TTG represent straightforward extensions of previous work: aspect recall [21], sub-topic recall [34], and the "nugget pyramid" approach from the TREC question answering evaluations [17]. Alternative metrics we had considered include those based on gain [7, 5] and the extension of mean average precision to graded relevance judgments [24]. The challenge with gain-based approaches is the complex parameterization necessary to fully instantiate a particular model; we lacked the empirical data to properly develop such a model. The graded extension to mean average precision is elegant, but our underlying user model is better captured by set-based metrics.

In total, 13 groups submitted 50 runs to the tweet timeline generation task at TREC 2014. Recognizing that systems make different choices with respect to balancing precision and recall, it is illustrative to visualize the tradeoffs in a scatter plot. Figure 2 shows precision vs. unweighted recall (left) and precision vs. weighted recall (right) for all runs. Iso-F_1 contours are plotted in blue; points on the same contour line have the same F_1 score, but with different precision/recall tradeoffs. Note that the effectiveness of individual runs is not relevant for the purposes of our meta-analysis, so we refer readers to the track overview for additional details [18].

4. ASSESSOR DIFFERENCES

Our first research question revolves around assessor differences in the semantic clustering task for tweet timeline generation and their impact on evaluation stability. For the TREC 2014 evaluation, we devoted the necessary resources to generate two independent sets of reference clusters, which we refer to as the "official" and "alternate" judgments. Note that the official judgments were simply the ones obtained first and used to report evaluation results at TREC; there is no implication that they are somehow more "authoritative".

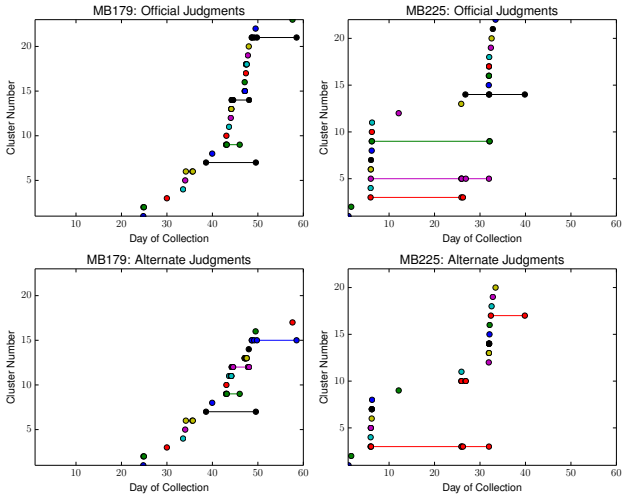

Figure 3: Visualization of the clusters for two topics.

4.1 Descriptive Characterization

We begin with a descriptive characterization of the semantic clusters generated by the assessors. In this and subsequent analyses, relevant and highly-relevant tweets are both considered "relevant". Each topic (55 in total) contains 194 relevant tweets on average: the official judgments averaged 89 clusters per topic, while the alternate judgments averaged 73 clusters per topic. These differences suggest that humans perform the semantic clustering task at different levels of granularity.

In Figure 3 we attempt to visualize these differences. In each plot, a point represents a tweet, and its x-axis position denotes the time when it was posted. Tweets that were assigned to the same cluster are at the same y-axis position and connected by horizontal lines. Thus, each horizontal "level" represents a semantic cluster, ordered by the first tweet; each cluster is assigned a different color for clarity. On the left, we show topic 179 "care of Iditarod dogs" and on the right, we show topic 225 "Barbara Walters, chicken pox". The top row shows the official judgments and the bottom row shows the alternate judgments. These topics were selected primarily for visual clarity: enough relevant tweets to show interesting clusters, but not too many relevant tweets as to be overly cluttered. From these visualizations, it is apparent that there are substantial assessor differences in the formation of the clusters. For example, the alternate assessor created fewer clusters for topic 179, and in topic 225, the alternate assessor did not seem to agree with two early clusters found by the official assessor.

To quantitatively characterize the differences between the official and alternate clusters, we computed the Adjusted Rand Index [23], which is a measure of similarity between two clusterings that is corrected for chance groupings (ranging from -1 to $+1$.). This metric has two other desirable properties: first, it is symmetric, which is appropriate since the official judgments aren't any more "correct" than the alternate judgments; second, it ignores permutations in that the similarity values don't depend on the "cluster labels" (which, in our case, are just arbitrary numeric identifiers). We computed the Adjusted Rand Index on a per topic basis ($N = 55$) between the official and alternate judgments and obtained a mean of 0.445, a median of 0.492, and a standard deviation of 0.225. As a point of reference, for topic 179 (left, Figure 3), the Adjusted Rand Index is 0.687, and for topic 225 (right, Figure 3), the Adjusted Rand Index is 0.305. We observe that the topics exhibit a wide range of similarity values, ranging from a minimum of 0.007 to a maximum of 0.977, which suggests that agreement is to a large extent dependent on the nature of the information need. However, it would be fair to say that there is substantial disagreement between assessors overall.

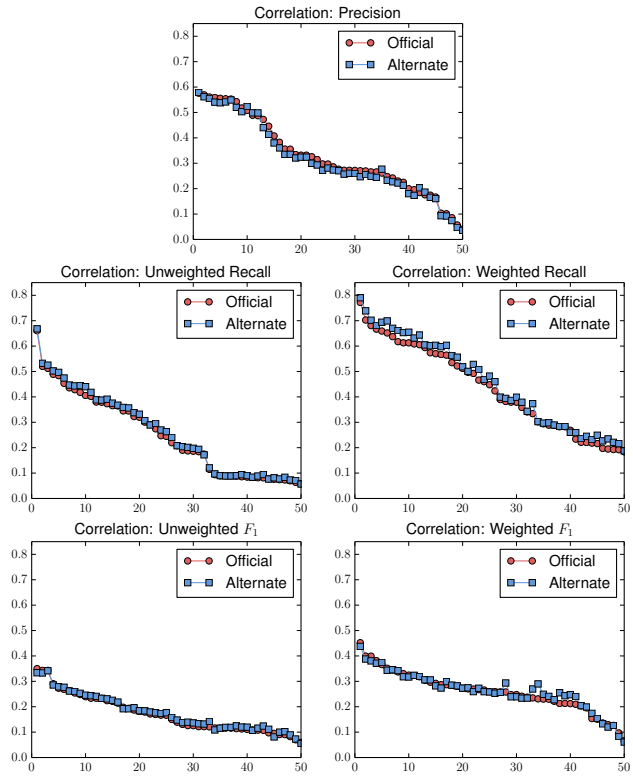

Figure 4: Comparison between scores based on the official judgments and the alternate judgments for various metrics. Runs are sorted by score based on the official judgments in descending order.

4.2 Stability Analysis

With the two independent sets of judgments, we can conduct a stability analysis of the evaluation to determine the extent to which assessor differences impact our ability to make system comparisons, i.e., that system X is more effective than system Y. An evaluation is considered stable if system rankings and pairwise system comparisons are insensitive to assessor differences. Here, we follow the well-established methodology of Voorhees [31], who examined assessor differences in document-level relevance judgments for ad hoc retrieval.

Figure 4 shows scores for all runs based on the official judgments and the alternate judgments for each of the five metrics. Results are sorted by scores based on the official judgments. We see that the rankings produced by both sets of judgments are highly correlated, with the exception of a few cases in weighted F_1. Furthermore, the absolute values of the metrics are also quite similar (particularly the unweighted metrics).

Metric	Rank Swaps	Kendall's τ
precision	30	0.951
unweighted recall	27	0.956
weighted recall	22	0.964
unweighted F_1	46	0.925
weighted F_1	90	0.853

Table 1: Count of rank swaps and Kendall's τ correlation based the official and alternate judgments for each metric.

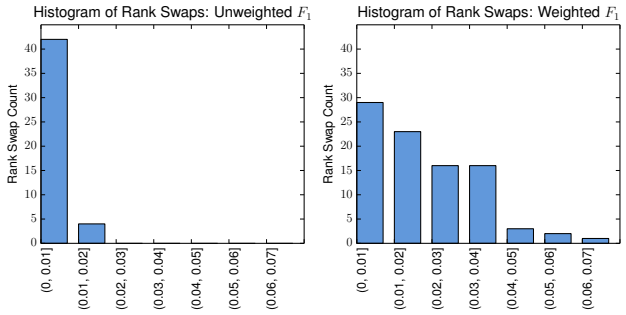

Figure 5: Histogram of rank swaps for unweighted F_1 and weighted F_1, binned by score differences.

In Table 1, we show the Kendall's τ correlation between rankings induced by the two different sets of judgments. These values are in the same range as those reported by Voorhees [31] for ad hoc retrieval, which is generally regarded by the IR community to be stable with respect to assessor differences in document-level relevance judgments. Table 1 also shows the number of rank swaps for each metric. A rank swap is a pairwise comparison where, according to one set of judgments, run A scores higher than run B, but according to the other set of judgments, run B scores higher than run A. There are a total of $(50 \times 49)/2 = 1225$ pairwise comparisons, so the numbers of rank swaps observed in Table 1 are quite small.

A tally of the rank swaps does not tell the complete story because rank swaps do not capture the magnitude of the score differences. In particular, we are less concerned with rank swaps in which the absolute score differences between the two conditions are small. Histograms of the rank swaps for unweighted F_1 and weighted F_1 binned by absolute score differences are shown in Figure 5. We see that, indeed, most of the rank swaps are small. For space considerations, we only show the histograms for these two metrics, which have the lowest Kendall's τ correlations. The histograms for the other metrics show even larger fractions of rank swaps where the score differences are small.

The conclusion from these analyses is fairly clear: assessor differences do not appear to impact the stability of the evaluation, at least in terms of the metrics we have examined. Although our findings are limited to tweet timeline generation, there is no reason to believe that these results would not carry over to other types of evaluations built around the notion of semantic clustering.

5. USER PREFERENCES

The second major research question we tackle in this paper concerns user preferences: when comparing TTG systems, do our metrics capture differences that are actually meaningful to users? Following previous work, we operationalize the notion of "meaningful" in terms of user preferences [2, 25]. If the evaluation tells us that system X is better than system Y, and a human can actually detect this difference (better than chance), we might reasonably claim that the improvement is meaningful from a user perspective.

5.1 Analysis Methodology

Unlike most studies discussed in Section 2, which focus on ad hoc retrieval (and variants), our analyses have a more complex setup because tweet timelines are variable in length. We are specifically interested in three distinct types of effectiveness differences:

- For two systems that exhibit roughly the same recall, how sensitive are users to differences in precision?
- For two systems that exhibit roughly the same precision, how sensitive are users to differences in recall?
- For two systems that exhibit similar precision-recall tradeoffs, how sensitive are users to differences in F_1?

Our general strategy was to sample system output from the submitted TREC 2014 TTG runs and to generate system comparisons on a per-topic basis for user preference assessment. Unlike some previous work that artificially manipulated system output to generate results of a particular level of effectiveness (e.g., [4, 30]), our procedure yields more realistic comparisons.

The sampling process proceeded as follows: First, we selected a set of 30 topics, biased toward "typical" topics that have neither too many nor too few relevant tweets. Let $x_i = r_i - \bar{r}$ be the difference between r_i, the number of relevant tweets for topic i, and $\bar{r}$, the median number of relevant tweets over all topics: the probability of "drawing" topic i is proportional to the density at x_i of a normal distribution with $\mu = 0$ and $\sigma = 20$ (chosen heuristically).

Next, we performed some filtering of the submitted runs: we discarded all runs that contained unjudged tweets to eliminate the effects of missing relevance judgments. We also discarded all runs longer than 41 tweets, which is the median length of submitted runs (after the first filter). A pilot study indicated that long runs are very difficult to judge, and this is roughly the point after which the judgments become too onerous to make.

After filtering, we randomly (i.e., uniformly) selected 20 "base" runs, and for each, sampled up to 20 different "comparison" runs per metric based on the following criteria:

- For what we call *precision sampling*, we selected runs that differed by less than 0.1 in recall, but more than 0.1 in precision. Previous work suggests that users have a hard time distinguishing small differences in effectiveness metrics, so we did not want to waste assessor effort.
- For what we call *recall sampling*, we selected runs that differed by less than 0.1 in precision, but more than 0.1 in recall. This is the opposite of precision sampling.
- For what we call F_1 *sampling*, we want the two comparison runs to make approximately the same tradeoff in terms of precision and recall. We selected runs where the value of $T = |P - R|$ for the base run is within 0.1 of the T for the comparison run, as long as the difference in *either* precision or recall between the two runs exceeds 0.1.

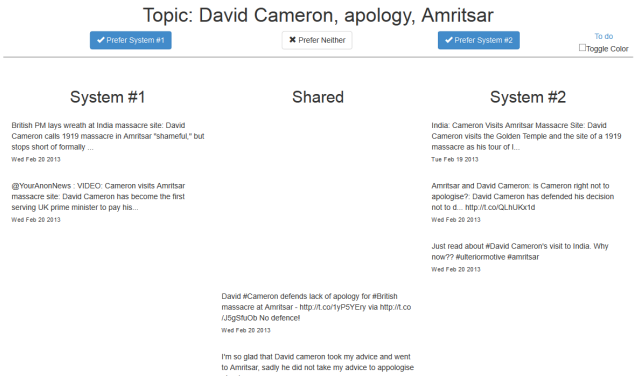

Figure 6: A screenshot of the web-based assessment interface for eliciting preference judgments. System outputs are presented in the left and right columns, with shared content in the middle column.

Note that all sampling was performed on the cluster-based TTG metrics, i.e., taking into account redundancy. We used the unweighted variants of the metrics for simplicity. Finally, we discarded all pairs where the two sampled runs differ in length by more than 20 tweets. Based on the pilot study, we found that assessors struggled to compare runs that differed significantly in length.

Each trial of the above sampling procedure yields a large number of comparisons, from which we further sampled 180 pairs comprised of roughly an equal number of pairs from precision, recall, and F_1 sampling. This represents a single batch that we gave to assessors to judge; assessors could request additional batches if they desired after completing a batch. The first batch for each assessor contained a shared set of 30 comparisons (i.e., all assessors judged the same pairs), but for the remaining pairs in the first batch (and all subsequent batches), each assessor worked on unique pairs, which we ensured in the batch preparation process.

For eliciting user preference judgments, we designed a web-based assessment interface (see screenshot in Figure 6). The comparison screen displays the sampled TTG outputs for a given topic side by side: System #1 on the left and System #2 on the right. Sampled pairs were mapped to the two positions randomly to avoid introducing any systematic biases. The topic is shown at the top. The output of each system is ordered chronologically, and each tweet is displayed with its timestamp. To help the assessor compare the two timelines, tweets returned by both systems are displayed in the middle "Shared" column. This means that an output might have gaps, indicating that one system returned more tweets than the other prior to the shared tweets (as is the case in Figure 6). We converged on this design after trying various display alternatives in a pilot study—we learned that assessors wanted an easy way to compare the two systems in terms of the tweets they *both* returned, so we designed the interface to facilitate this comparison.

In the assessment instructions, we asked the assessors which of the two system outputs they thought was better. They were reminded to evaluate system outputs as timelines, rather than as ranked lists. To be consistent with the clustering task, assessors were told to evaluate tweets based solely on their contents, ignoring any links they may contain. The assessor could click "Prefer System #1" or "Prefer System #2" above each of the two conditions, to indicate a preference, or a button labeled "Prefer Neither" in the middle if the assessor could not decide.

We avoided prescriptively dictating what it meant for a timeline to be "good". Specifically, the instructions made no reference to precision, recall, or redundancy. This is an important feature in our evaluation design to mitigate the effects of demand characteristics.[3] However, we tried to help the assessors frame their task with a few neutral reminders:

- We are asking you about relative quality. Even if you feel that both systems do a bad job, do your best to determine which does a better job.
- The tweets are presented in chronological order (earliest first), so you should keep in mind the timeframe covered by each result.
- The timelines are displayed with a "shared" column in the middle. This column contains all tweets that were returned by both systems, and should help you more quickly determine how similar the results are. Since the timelines are aligned by their shared tweets, this middle column will also help you compare the results according to chronological blocks.

Assessors were first trained in the laboratory. They were given an introduction to the task and an explanation of the web interface, which included time spent with practice topics that were discarded for analysis. After this training session, the assessors could proceed at their own pace anywhere they had an internet connection. The server that hosted the web assessment interface recorded all interactions: both the preference judgments and their timestamps.

Evaluation proceeded concurrently at UMD and UIUC. Judgments at UMD were performed by two male computer science graduate students. One of these assessors completed two batches of comparisons, while the other completed one. The assessors at UIUC were two graduate students in library and information science (one male, one female) and one former library and information science graduate student (male) who was working nearby. One of the assessors completed two batches, while the remaining assessors completed one each. These assessors were not the same as the annotators who created the clusters (at both sites).

5.2 Results

After the assessments were completed, the user preferences were correlated against the preferences implied by the sampled metric, using Cohen's κ as the agreement metric. Cohen's κ ranges from -1 to $+1$, where 0 indicates that any agreement is attributable to chance. To compute κ, we first discarded "prefer neither" judgments. In the precision sampling case, we correlated the human preferences against the implied preference based on the precision scores. In the recall sampling case, we correlated the human preferences against the implied preference based on both unweighted recall and weighted recall. In the F_1 sampling case, we correlated the human preferences against the implied preference based on unweighted F_1 and weighted F_1.

Figure 7 shows these correlations binned by differences in the sampled metric, aggregated across all assessors. The

[3] An experimental artifact where participants form an interpretation of the experiment's purpose and unconsciously change their behavior to fit that interpretation.

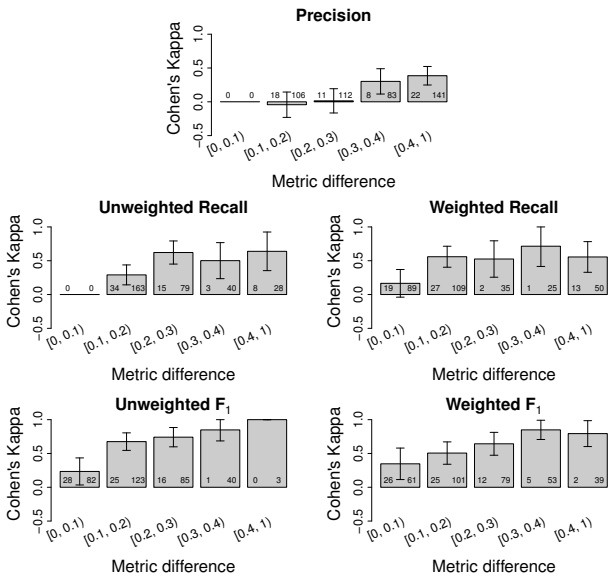

Figure 7: Cohen's κ for each metric binned by differences in the sampled metric. Error bars show 95% confidence intervals. The numbers in the bottom left of each bar show the number of "prefer neither" while those in the bottom right show the number of preference judgments in that condition.

error bars show the 95% confidence intervals, computed following Fleiss et al. [11]. The numbers in the bottom of each bar show the number of "prefer neither" (on the left) and the number of preference judgments, i.e., users preferred one system over the other (on the right). Note that by the design of the sampling procedure, no comparisons fell into the $[0, 0.1)$ bucket for precision and unweighted recall, although some samples did fall into that bucket for the other metrics.

In Figure 8, we show agreement broken down by each individual assessor in a "small multiples" visualization scheme. Each row shows a particular metric, and each column shows the agreement for an individual assessor for that metric. For clarity, the bar charts are shown without labels, but they are organized in exactly the same manner as in Figure 7. Note that confidence intervals here are much larger because fewer samples fall into each bin.

In both Figures 7 and 8, the bars are arranged by putative difficulty of the preference judgments, from left to right. Since there were generally fewer samples in which the metric differed by more than 0.4, all difference greater than 0.4 were placed in the same bin. The leftmost bar represents those runs where the metric difference is less than 0.1, for which we would expect humans to have the most difficulty distinguishing differences in output quality. Similarly, the rightmost bar should represent the "easiest" comparisons, in that the metric differences are the largest. We can intuitively think of these charts as quantifying how "sensitive" users are to differences in the underlying system-oriented metrics. Put another way, they tell us how much "better" or "worse" a system would have to be in order for users to notice.

Overall, we do see a general trend of the agreement increasing from left to right, both at the aggregate and at the individual level. At the aggregate level, we see that there tend to be more "prefer neither" judgments in cases where the metric difference is small, which makes sense. However, there are clearly individual differences that buck the overall trend, as shown in Figure 8. For example, assessor 4 (fourth column from the left) does not appear to be sensitive to precision at all—agreement remains near and sometimes worse than chance, even for very large differences in precision.

κ	Strength of Agreement
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

Table 2: Interpretations of κ from [16].

Metric	F-value	p-value
precision	$F(3, 16) = 3.397$	0.0437
unweighted recall	$F(3, 16) = 1.749$	0.1970
weighted recall	$F(4, 20) = 1.475$	0.2470
unweighted F_1	$F(3, 16) = 5.093$	0.0115
weighted F_1	$F(4, 20) = 1.322$	0.2960

Table 3: ANOVA results for each metric. Metrics displayed in bold show a statistically significant difference ($p < 0.05$) in mean κ across bins.

One of the downsides of κ is that it lacks an intuitive interpretation. To address this, Landis and Koch [16] proposed a guide, which we replicate in Table 2. Under this rubric, the minimum κ values we observe per metric range from "poor" (for precision) to "fair" (for weighted F_1), whereas the maximum κ values we observe for each metric range from "fair" (for precision) to "almost perfect" (for unweighted and weighted F_1). The differences between maximum and minimum κ values per metric support the apparent relationship between metric differences and annotator preferences.

It is worth emphasizing here that all our metrics already take into account redundancy (i.e., they are cluster-based). Since we *did not* explicitly ask the assessors to consider redundancy in the assessment instructions, these results show that humans are nevertheless sensitive to the redundancy removal performed by the systems.

To add statistical rigor to our analyses about the relationship between agreement and metric differences, we performed analysis of variance (ANOVA) comparing the κ values across bins. That is, we treat the κ of each individual assessor for a particular bin as a point estimate of the "true" κ for that bin (magnitude of metric difference), and want to know if the mean κ values across the bins are significantly different. The results are shown in Table 3. We found statistically significant differences in mean κ across bins for precision and unweighted F_1 at $p < 0.05$. From this, we can conclude that increases in metric differences are associated with better agreement for those metrics.

A surprising result is the lack of statistical significance for both unweighted and weighted recall. What does this mean? The recall bar charts in Figure 7 show that differences in recall between pairs of runs *are* detectable by annotators; however, the *magnitude* of the difference does not appear to affect the agreement, i.e., the κ values are not significantly different across the bins. We believe that this is due to the nature of evaluating recall, which requires knowledge of the

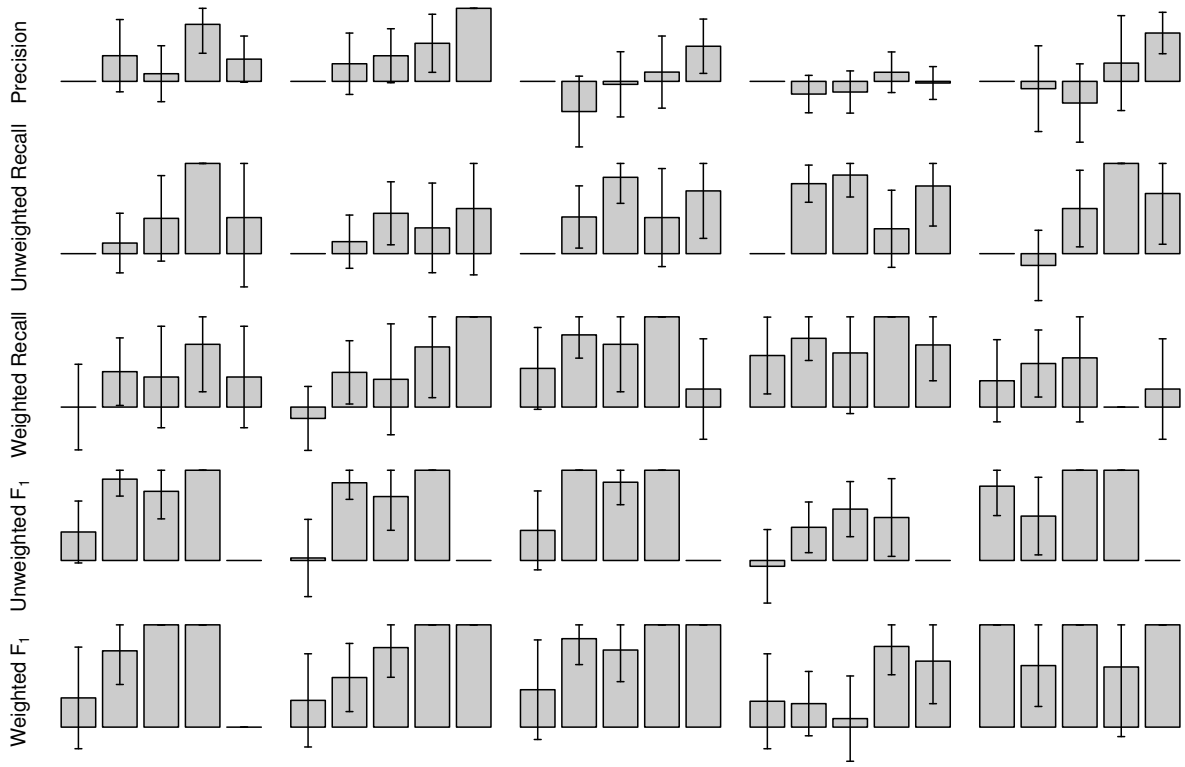

Figure 8: Cohen's κ for each metric (row), for each individual assessor (column) arranged in "small multiples". Each bar chart is organized in the same manner as those in Figure 7.

complete relevance judgments. When comparing two runs, it is relatively easy to detect recall differences (i.e., if one system returned a relevant tweet that was absent from the other system), but without access to all the relevance judgments, it is difficult to tell *how much* better one is than the other (i.e., the marginal recall increase from returning one more relevant tweet is dependent on the number of relevant tweets for that topic). Note that in contrast, for precision, annotators *do* appear capable of detecting the magnitude of the difference. This is again unsurprising since precision can be assessed directly from the displayed tweets.

We were surprised at the high p-value for weighted F_1, particularly in comparison to unweighted F_1. This result indicates that although assessors consistently achieved better than chance agreement, they either had difficulty assessing the importance of individual tweets or our weighting scheme does not accurately capture users' notion of importance. The second possibility points to future work in developing more meaningful metrics for modeling importance.

Figure 7 shows another surprising result: we observe relatively low agreement for precision (compared to the other metrics). We believe that this may be an artifact of our sampling strategy, where we discarded overly verbose runs. This choice was made out of necessity, as our pilot study suggested that it was not feasible to ask assessors to compare timelines with, say, hundreds of tweets. The problem is this: while it is certainly possible to obtain low precision in a short timeline, the amount of "pain" associated with such low precision timelines is relatively low (bounded by the total number of returned tweets). However, if a timeline were an order of magnitude longer, say, the burden imposed by poor precision would be much heavier. In other words, our sampling procedure selected only those runs where differences in precision were "not a big deal" from the assessor's point of view because the timelines were relatively short.

We also analyzed the relationship between the time assessors spent on each comparison (reconstructed from our logs) and the observed agreement. This is shown in Figure 9, binned in intervals of 30 seconds for each metric. The numbers at the top of each bar show the number of "prefer neither" and preference judgments. "Prefer neither" judgments were not included in the agreement calculation. Missing bars indicate an undefined κ while dashes indicate a κ of zero. We see no clear relationship between κ and the amount of time it takes to make a judgment; this is somewhat surprising as we would have expected "easy judgments" to result in higher agreement.

Finally, we analyzed agreement on the set of 30 shared comparisons provided to each of the five annotators at the beginning of the first batch. We calculated Fleiss' κ [10], a variant of Cohen's κ intended for more than two raters, and found a value of 0.294 for this shared set. Under Landis and Koch's scheme, this represents only a fair level of agreement among annotators. This finding is noteworthy because it indicates that user preferences correlate with metric differences, *despite* the fact that humans only achieve "fair" agreement among themselves.

Summarizing these results, our user study suggests that user preferences do agree with system preferences as measured by the TTG metrics—that users are sensitive to differences in precision, recall, and F_1, although to different degrees. We can therefore conclude that system-oriented metrics are meaningful in being able to capture aspects of what users care about in timeline summaries.

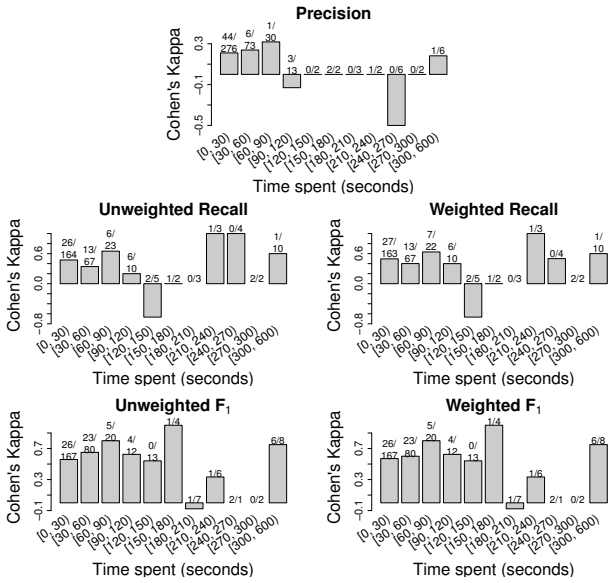

Figure 9: Cohen's κ binned by time spent on each comparison (at 30 second intervals). Numbers at the top of each bar show the number of "prefer neither" and preference judgments. Missing bars indicate an undefined κ while dashes indicate a κ of zero.

6. CONCLUSIONS

As primarily an empirical discipline, progress in information retrieval is built on system comparisons. This paper explores assessor differences and user preferences in the context of the tweet timeline generation task in the TREC 2014 Microblog track. Analyses show that the evaluation methodology appears to be sound, which strengthens our confidence in existing and future results. Although our findings are limited to this particular task, we believe that lessons learned could generalize to other "cluster-based" evaluations, informing related tasks such as question answering and temporal summarization.

7. ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under IIS-1217279 and IIS-1218043. Any opinions, findings, conclusions, or recommendations expressed are those of the authors and do not necessarily reflect the views of the sponsor. We are grateful to Ellen Voorhees and the assessors at NIST for making TREC possible.

8. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. *WSDM*, 2009.

[2] A. Al-Maskari, M. Sanderson, P. Clough, and E. Airio. The good and the bad system: Does the test collection predict users' effectiveness? *SIGIR*, 2008.

[3] J. Allan. *Topic Detection and Tracking: Event-Based Information Organization*. Kluwer, 2002.

[4] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? User effectiveness as a function of retrieval accuracy. *SIGIR*, 2005.

[5] J. Aslam, M. Ekstrand-Abueg, V. Pavlu, R. McCreadie, F. Diaz, and T. Sakai. TREC 2014 temporal summarization track overview. *TREC*, 2014.

[6] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. de Vries, and E. Yilmaz. Relevance assessment: Are judges exchangeable and does it matter? *SIGIR*, 2008.

[7] C. Clarke, M. Kolla, G. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. *SIGIR*, 2008.

[8] H. Dang and J. Lin. Different structures for evaluating answers to complex questions: Pyramids won't topple, and neither will human assessors. *ACL*, 2007.

[9] H. Dang, J. Lin, and D. Kelly. Overview of the TREC 2006 question answering track. *TREC*, 2006.

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

[11] J. Fleiss, J. Cohen, and B. Everitt. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72(5):323–327, 1969.

[12] Q. Guo, F. Diaz, and E. Yom-Tov. Updating users about time critical events. *ECIR*, 2013.

[13] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson. Do batch and user evaluations give the same results? *SIGIR*, 2000.

[14] S. Huffman and M. Hochster. How well does result relevance predict session satisfaction? *SIGIR*, 2007.

[15] J. Hurlock and M. Wilson. Searching Twitter: Separating the tweet from the chaff. *ICWSM*, 2011.

[16] J. Landis and G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[17] J. Lin and D. Demner-Fushman. Will pyramids built of nuggets topple over? *NAACL*, 2006.

[18] J. Lin, M. Efron, Y. Wang, and G. Sherman. Overview of the TREC-2014 Microblog track. *TREC*, 2004.

[19] J. Lin and M. Smucker. How do users find things with PubMed? Towards automatic utility evaluation with user simulations. *SIGIR*, 2008.

[20] J. Lin and P. Zhang. Deconstructing nuggets: The stability and reliability of complex question answering evaluation. *SIGIR*, 2007.

[21] P. Over. TREC-6 interactive report. *TREC*, 1997.

[22] V. Pavlu, S. Rajput, P. Golbus, and J. Aslam. IR system evaluation using nugget-based test collections. *WSDM*, 2012.

[23] W. Rand. Objective criteria for the evaluation of clustering methods. *JASA*, 66(336):846–850, 1971.

[24] S. Robertson, E. Kanoulas, and E. Yilmaz. Extending average precision to graded relevance judgments. *SIGIR*, 2010.

[25] M. Sanderson, M. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? *SIGIR*, 2010.

[26] C. Smith and P. Kantor. User adaptation: Good results from poor systems. *SIGIR*, 2008.

[27] M. Smucker and C. Jethani. Human performance and retrieval precision revisited. *SIGIR*, 2010.

[28] K. Tao, C. Hauff, and G.-J. Houben. Building a microblog corpus for search result diversification. *AIRS*, 2013.

[29] A. Turpin and W. R. Hersh. Why batch and user evaluations do not give the same results. *SIGIR*, 2001.

[30] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. *SIGIR*, 2006.

[31] E. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *SIGIR*, 1998.

[32] E. Voorhees. Overview of the TREC 2004 question answering track. *TREC*, 2004.

[33] C. Wayne. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. *LREC*, 2000.

[34] C. Zhai, W. Cohen, and J. Lafferty. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. *SIGIR*, 2003.