Track Proposal: TREC 2017 Real-Time Summarization (RTS) Track

Jimmy Lin, Adam Roegiest, and Luchen Tan (University of Waterloo)
Richard McCreadie (University of Glasgow)

There is increasing interest in systems that can automatically monitor streams of social media posts, such as tweets on Twitter, with the aim of providing concise updates to users about topics they care about. We might think of these topics as "interest profiles", specifying users' prospective information needs, i.e., what they want to receive updates about in the future. For example, a user might be interested in poll results for the 2016 U.S. presidential elections and wishes to be notified whenever new results are published.

At TREC 2014 and 2015, the Microblog (MB) track and the Temporal Summarization (TS) tracks examined how to build systems that produce real-time updates about specified topics given streaming data, namely tweets and news articles, respectively. For TREC 2016, the Real-Time Summarization (RTS) track was created, which represented a merger of the MB and TS tracks. The creation of the RTS track was designed to leverage synergies between the two tracks in exploring prospective information needs over document streams containing novel and evolving information, while also breaking new ground in terms of evaluation methodology via the real-time judging of updates as they are produced.

The TREC 2016 RTS track was successful, with 89 runs from 20 participating groups across two evaluation scenarios. For TREC 2017, we propose to continue running the RTS track. From the participant perspective, this will provide a higher quality test collection with a broader variety of user interest profiles and associated content, thereby enabling better characterizations of real-time summarization system quality. Moreover, by participant request, we propose to introduce the dynamic expansion of interest profiles over the assessment period, enabling the addition of new profiles related to events that were not predicted at the start of the assessment period. Furthermore, for the second year of the track, having prepared the groundwork for generating real-time assessments in 2016, we are considering adding live system feedback to participants on their updates. This would open new doors for experimentation with active learning systems, enabling us to expand the scope of techniques that can be brought to bear on the update summarization problem.

TREC 2016 RTS Track: Experimental Design

For systems that monitor social media streams with respect to users' information needs, we can imagine two methods for disseminating updates:

Scenario A: Push notifications. As soon as the system identifies a relevant post, it is immediately sent to the user's mobile phone via a push notification. At a high level, push notifications should be relevant (on topic), timely (provide updates as soon after the actual event occurrence as possible), and novel (users should not be pushed multiple notifications that say the same thing).

Scenario B: Email digest. Alternatively, a user might want to receive a daily email digest that summarizes "what happened" that day with respect to the interest profiles. At a high

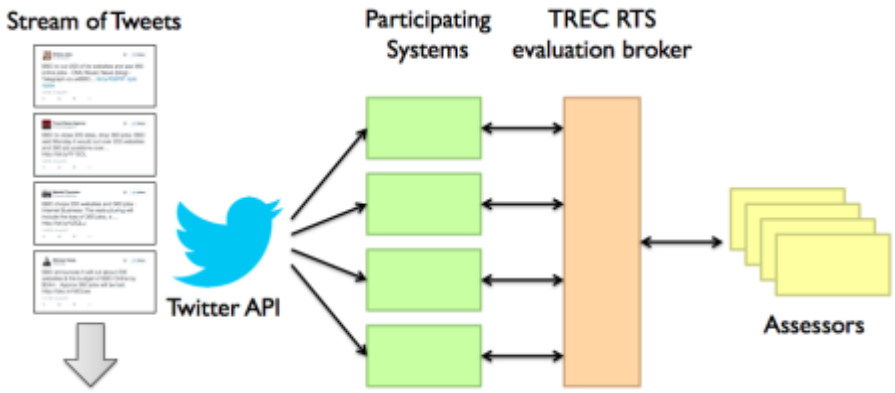
level, these results should be relevant and novel; timeliness is not as important, provided that the tweets were all posted on the previous day.

The RTS track in TREC 2016 operationalized both scenarios.

For more details, the track homepage is located at http://trecrts.github.io/ and the official track guidelines are located at http://trecrts.github.io/TREC2016-RTS-guidelines.html.

The evaluation occurred from August 2, 2016 00:00:00 UTC to August 11, 2016 23:59:59 UTC. During this time, all participating systems "listened" to the Twitter sample stream using the Twitter streaming API and performed the evaluation tasks (either scenario A or scenario B) in real time. Prior to the beginning of the evaluation period, systems were provided a list of "interest profiles" (similar to topics in ad hoc retrieval) representing users' information needs.

Scenario A systems were evaluated in two different ways: The first is live user-in-the-loop assessments, as illustrated in the following figure:



That is, content identified as relevant by a system based on the user's interest profile in real time were pushed to the TREC RTS evaluation broker (via a REST API). These notifications were then immediately delivered to the mobile phones of a group of assessors. This evaluation methodology was introduced for the first time in TREC 2016 and promises a number of significant advantages over traditional post hoc batch evaluations because it is able to capture live user assessments. However, scenario A systems were also evaluated with a traditional post hoc batch evaluation methodology (derived from previous TREC Microblog evaluations) to facilitate post hoc meta-evaluations.

Scenario B systems were only evaluated with a traditional post hoc batch evaluation methodology (derived from previous TREC Microblog evaluations).

TREC 2016 RTS Track: Evaluation Status

For scenario A, we received 49 runs from 20 groups (each were allowed a max of three runs). In total, these runs pushed a total of 161,729 tweets during the evaluation period. After de-duplication within topics, the RTS broker received 95,115 unique tweets.

We recruited 18 students from the University of Waterloo to serve as the mobile assessors, 13 of whom ultimately provided judgments during the evaluation period. In total, we received 12,115 judgments over the evaluation period, with a minimum of 28 and a maximum of 3,791 by an individual assessor.

For scenario B, we received 40 runs from 15 groups.

Post-hoc assessment of scenario A and scenario B runs are currently in progress. We have begun to analyze the judgments received from the live human assessors for scenario A.

TREC 2017 RTS Track: Refinements

Since scenario A systems were evaluated using both live human assessors as well as a traditional post hoc batch methodology, we can conduct a meta-evaluation to examine the reliability of this new user-in-the-loop approach. These analyses, however, cannot begin until we receive the batch evaluation results, which we expect shortly. Refinements for the RTS track in TREC 2017 will be guided by these analyses.

There are, however, a number of observations we have made about the execution of the track this year that serve as the starting point of proposed improvements for next year:

Refine the assessment app. Despite substantial effort, we were not able to successfully deploy a working assessment app for iOS in time. Therefore, only the Android app was available, which limited the size of our assessor pool. Furthermore, there were many display bugs that were encountered by the assessors during the evaluation period, which stemmed from the diversity of devices used by the assessors (and was difficult to anticipate in advance).

For TREC 2017, we will build on our experiences to improve the assessor experience. Plans include rewriting a new assessment app for Android and iOS, fixing all reported bugs from this year, and taking advantage of our existing evaluation infrastructure to perform test runs in advance of the official evaluation period.

Refine the tweet routing strategy. When the RTS evaluation broker receives a tweet pushed by a system, it must be routed to the assessors. The simplest routing strategy – and what we implemented this year – is to forward the tweet to all assessors who have subscribed to that particular interest profile. The advantage of this approach is that we might potentially receive multiple judgments on the same tweet, allowing us to compute inter-assessor agreement. However, such a strategy may not be the most effective use of scarce assessor resources. We can imagine alternative strategies such as round-robin distribution, or even more sophisticated techniques that take into account the users' response times. For TREC 2017, we will explore different tweet routing strategies based on data we have gathered this year.

Support user-supplied interest profiles. The interest profiles (i.e., topics) this year were developed prior to the evaluation period, and thus needed to anticipate future events. It would have been preferable to solicit interest profiles from the assessors themselves (and to introduce new interest profiles during the evaluation period) so they could judge the relevance of tweets with respect to information needs that were their own. Although we contemplated implementing such a mechanism for this year, we abandoned the efforts early

on because of the system complexities involved. For TREC 2017, we believe that is both possible and practical to introduce such a mechanism.

Live Feedback: For the 2016 edition of the track, participant systems submitted updates in real time as they processed the live tweet stream. These updates were pushed to assessors on their mobile phones for assessment with as little latency as possible, resulting in close-to real-time generation of relevance judgments. A natural extension of this methodology would be to make these judgments immediately available to the participating systems, such that they can use the feedback to learn on the fly what updates users are interested in. For 2017, we are considering the addition of live feedback to scenario A, although we are hesitant to increase the complexity of the task at this early stage in the track's lifetime.