On the Evaluation of Tweet Timeline Generation Task

Walid Magdy[1(✉)], Tamer Elsayed[2], and Maram Hasanain[2]

[1] Qatar Computing Research Institute, HBKU, Doha, Qatar
wmagdy@qf.org.qa
[2] Computer Science and Engineering Department, Qatar University, Doha, Qatar
{telsayed,maram.hasanain}@qu.edu.qa

Abstract. Tweet Timeline Generation (TTG) task aims to generate a timeline of *relevant* but *novel* tweets that summarizes the development of a given topic. A typical TTG system first retrieves tweets then detects novel tweets among them to form a timeline. In this paper, we examine the dependency of TTG on retrieval quality, and its effect on having biased evaluation. Our study showed a considerable dependency, however, ranking systems is not highly affected if a common retrieval run is used.

1 Introduction

With the enormous volume of tweets posted daily and the associated redundancy and noise in such vibrant information sharing medium, a user can find it difficult to get updates about a topic or an event of interest. The Tweet Timeline Generation (TTG) task was recently introduced at TREC-2014 microblog track to tackle this problem. TTG aims at generating a timeline of *relevant* and *novel* tweets that summarizes the development of a topic over time [5].

In the TREC task, a TTG system is evaluated using variants of F_1 measure that combine precision and recall of the generated timeline against a gold standard of clusters of semantically-similar tweets. Different TTG approaches were presented in TREC-2014 [5] and afterwards [2,4]: *almost all* rely on an initial step of retrieval of a ranked list of potentially-relevant tweets, followed by applying novelty detection and duplicate removal techniques to generate the timeline [5]. In such design, the quality of generated timeline naturally relies on that of the initially-retrieved list. There is a major concern that the evaluation metrics do not fairly rank TTG systems since they start from *different* retrieved ranked lists. An effective TTG system that is fed low quality list may achieve lower performance compared to another low quality TTG system fed a high quality list; current TTG evaluation metrics lacks the ability to evaluate TTG independently from the retrieval effectiveness. This creates an evaluation challenge, especially for future approaches that use different retrieval models.

In this work, we examine the bias of TTG evaluation methodology introduced in the track [1]. We first empirically measure the dependency of TTG

N. Ferro et al. (Eds.): ECIR 2016, LNCS 9626, pp. 648–653, 2016.
DOI: 10.1007/978-3-319-30671-1_48

performance on retrieval quality, then examine the validity of using a single input retrieval list for ranking different TTG systems, and the consistency of ranking when using several retrieval lists with varying qualities. We ran experiments using 13 different ad-hoc retrieval runs and 8 TTG systems participated in TREC-2014. Our study shows considerable dependency of TTG systems performance on retrieval quality. Nonetheless, we found that using a single ad-hoc run for ranking different TTG systems could lead to a less-biased and stable ranking of TTG systems, regardless of which retrieval run is used. When a common retrieval run is not available, it is important to consider the final performance of the TTG system in the context of the quality of the used retrieval run.

2 Experimental Setup

A set of 55 queries and corresponding relevance judgments were provided by TREC [5]. For each query, a set of semantic clusters were identified; each consists of tweets relevant to an aspect of the topic but substantially similar to each other.

Precision, recall, and F_1 measures over the semantic clusters were used for evaluation. Precision (**P**) is defined as the proportion of tweets returned by a TTG system representing distinct semantic clusters. Recall (**R**) is defined as the proportion of the total semantic clusters that are covered by the returned tweets. Weighted Recall (**wR**) is measured similarly but weighs each covered semantic cluster by the sum of relevance grades[1] of its tweets. F_1 combines P and R, while wF_1 combines P and wR. Each of those measures is first computed over the returned timeline of *each* query and then averaged over all queries.

In our experiments, we used 12 officially-submitted ad-hoc runs by 3 of the top 4 participated groups in TREC-2014 TTG task [3,6,9]. Additionally, we used a baseline run directly provided by TREC search API [5]. This concludes a total of 13 ad-hoc runs for our study, denoted by the set $\mathrm{A} = \{a_1, a_2, \ldots, a_{13}\}$. The retrieval approaches used by those runs are mainly five: (1) direct search by TREC API (D), (2) using query expansion (QE), (3) using QE that utilizes the links in tweets (QE+Web), (4) using QE then learning to rank (QE+L2R), and (5) using relevance modeling (RM). Table 1 presents all ad-hoc runs and their retrieval performance.

We also used 8 different TTG systems (of two TREC participants) [3,6], denoted by $\mathrm{T} = \{t_1, t_2, \ldots, t_8\}$. Their approaches are summarized as follows:

- t_1 to t_4 applied 1NN-clustering (using modified versions of Jaccard similarity) on the retrieved tweets [6] and generated timelines using different retrieval depths, which made their performance results significantly different [5,6].
- t_5 is a simple TTG system that just returns the retrieved tweets after removing exact duplicates.
- t_6 to t_8 applied an incremental clustering approach that treats the retrieved tweets, sorted by their retrieval scores, as a stream and clusters each tweet based on cosine similarity to the centroids of existing clusters. They also used different number of top retrieved tweets and different similarity thresholds, and considered the top-scoring tweet in each cluster as its centroid [3].

[1] 1 for a relevant tweet and 2 for a highly-relevant tweet.

Table 1. Retrieval performance of ad-hoc runs.

Ad-hoc	MAP	P@30	P@100	R-Prec	Approach
a_1	0.477	0.669	0.500	0.501	QE+web
a_2	0.482	0.698	0.500	0.501	QE+L2R
a_3	0.464	0.668	0.491	0.498	QE
a_4	0.470	0.699	0.491	0.498	QE+L2R
a_5	0.490	0.670	0.505	0.508	QE
a_6	0.466	0.644	0.479	0.496	QE
a_7	0.406	0.647	0.461	0.445	QE
a_8	0.445	0.627	0.509	0.486	RM
a_9	0.385	0.624	0.473	0.436	RM
a_{10}	0.485	0.673	0.517	0.499	QE+web
a_{11}	0.497	0.681	0.518	0.512	QE+L2R
a_{12}	0.571	0.712	0.545	0.566	QE+L2R
a_{13}	0.398	0.646	0.468	0.439	D

Table 2. TTG systems performance with a_{13}.

TTG	R	wR	P	F_1	wF_1
$t_1(a_{13})$	0.342	0.535	0.320	0.245	0.330
$t_2(a_{13})$	0.260	0.463	0.411	0.241	0.371
$t_3(a_{13})$	0.159	0.261	0.364	0.153	0.226
$t_4(a_{13})$	0.137	0.261	0.444	0.150	0.255
$t_5(a_{13})$	0.353	0.552	0.263	0.231	0.297
$t_6(a_{13})$	0.315	0.511	0.354	0.252	0.350
$t_7(a_{13})$	0.191	0.365	0.484	0.215	0.355
$t_8(a_{13})$	0.334	0.537	0.311	0.246	0.328

Table 2 presents the performance of the 8 TTG systems when applied to a_{13}, which was selected as a sample to illustrate the quality of each TTG system. As shown, the quality of the 8 TTG systems varies significantly. In fact, by applying significance test on wF_1 using two-tailed t-test with α of 0.05, we found that all TTG system pairs but 6 were statistically significantly different.

Combinations of ad-hoc runs and TTG systems created a list of **104** TTG runs that we used to study the bias of the task evaluation. We aim to show whether the evaluation methodology used in the TREC microblog track is biased towards retrieval quality, and if there is a way to reduce possible bias.

To measure bias and dependency of TTG on the quality of the used ad-hoc runs, we use Kendall tau correlation (τ) and AP correlation (τ_{AP}) [10]. τ_{AP} is used besides τ since it is more sensitive to errors at higher ranks [10].

3 Dependency of TTG Performance on Retrieval Results

3.1 Correlation Between TTG Scores and Retrieval Scores

In this section, we try to answer the following research question: *"If we tried one TTG system with different ad-hoc runs, will the quality ranking of the resulting TTG timelines be correlated with the quality ranking of the ad-hoc runs?"*.

To answer this question, we compared the ranking of the ad-hoc runs (using retrieval scores) to the ranking of the resulting timelines from the same TTG system (using TTG scores). We repeated the process over each TTG system, and averaged the correlations as follows:

$$\sigma^* = \frac{1}{|T|} \sum_{t \in T} \sigma(\{S_r(a)|_{a \in A}\}, \{S_t(t(a))|_{a \in A}\}) \quad (1)$$

where σ^* is the average correlation, σ is τ or τ_{AP} correlation, $\{S_r(a)|_{a \in A}\}$ are the retrieval scores of the 13 ad-hoc runs, and $\{S_t(t(a))|_{a \in A}\}$ are the TTG scores of their corresponding timelines using the TTG system t.

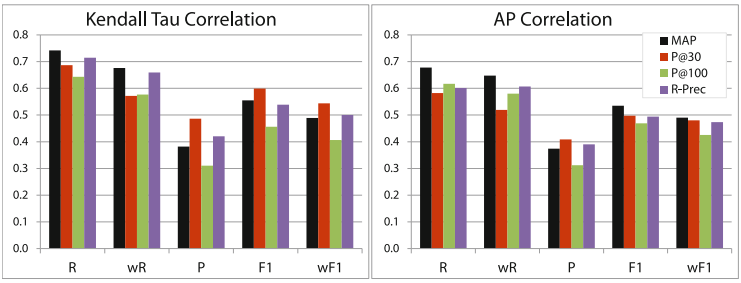

Fig. 1. τ and τ_{AP} between ad-hoc runs and their corresponding TTG timelines, averaged over TTG systems.

Figure 1 reports the *average* τ and τ_{AP} correlations using different retrieval and TTG performance metrics. As shown, there is always a positive correlation between the quality rankings of ad-hoc runs and the TTG timelines. Considering the main metrics for evaluating retrieval (MAP) and for evaluating TTG (wF_1), the correlation values are 0.49 for both τ and τ_{AP}. This indicates a considerable correlation, but it is not very strong as expected.

3.2 Correlation over TTG Scores

Since measuring correlation between systems ranking using measures of two different tasks may be sub-optimal, we continue to test the dependency of TTG output on the ad-hoc runs quality, but using TTG evaluation measures only.

Here we answer the following research question: *"If we tried a TTG system t_i with different ad-hoc runs, and we repeated that with another TTG system t_j, will the quality ranking of the resulting timelines of t_i be correlated with the quality ranking of the resulting timelines of t_j?"*.

Correlation is computed between the ranking of resulting timelines from TTG system t_i using different ad-hoc runs and the corresponding timelines from TTG system t_j. We apply this over the 8 TTG systems, creating a set of 28 pairwise comparisons. The average correlation is then computed as follows:

$$\sigma^* = \frac{2}{|T|(|T|-1)} \sum_{i=1}^{|T|} \sum_{j=i+1}^{|T|} \sigma(\{S_t(t_i(a))|_{a \in A}\}, \{S_t(t_j(a))|_{a \in A}\}) \quad (2)$$

Table 3 reports the average τ and τ_{AP} correlations among all pairs of TTG systems. Achieved correlation scores align with the same ones in Fig. 1, but with slightly higher values. This also supports the finding that TTG system performance depends, to some extent, on the quality of input ad-hoc runs. This observation suggests that using different ad-hoc runs with different TTG systems makes it unlikely to have unbiased evaluation for the TTG systems, since the output of TTG systems, in general, depends on the quality of the retrieval run.

Table 3. Average correlation between rankings of pairs of TTG systems when using all ad-hoc runs.

σ^*	R	wR	P	F_1	wF_1
avg. τ	0.76	0.57	0.57	0.68	0.56
avg. τ_{AP}	0.71	0.52	0.50	0.63	0.51

Table 4. Average correlation between rankings resulted from pairs of ad-hoc runs when used for all TTG systems.

σ^*	R	wR	P	F_1	wF_1
avg. τ	0.96	0.97	0.93	0.86	0.85
avg. τ_{AP}	0.92	0.93	0.81	0.72	0.76

4 Performance Stability Across Multiple Ad-Hoc Runs

In this section, we study stability of performance of a TTG system using different ad-hoc runs. For example, we examine if the best-performing TTG system using one ad-hoc run would continue to be the best with other ad-hoc runs.

We specifically investigate the following research question: *"If we used an ad-hoc run a_i with different TTG systems, and we repeated that with another ad-hoc run a_j, will the quality ranking of the resulting timelines using a_i be correlated with the quality ranking of the resulting timelines using a_j?"*

We compute correlation between quality ranking of the resulting timelines of the 8 TTG systems when using ad-hoc run a_i and the corresponding ranking when using run a_j. We apply that over all pairs of the 13 ad-hoc runs, creating a set of 78 pairwise comparisons. Average correlation is computed as follows:

$$\sigma^* = \frac{2}{|A|(|A|-1)} \sum_{i=1}^{|A|} \sum_{j=i+1}^{|A|} \sigma(\{S_t(t(a_i))|_{t \in T}\}, \{S_t(t(a_j))|_{t \in T}\}) \quad (3)$$

Table 4 reports the average τ and τ_{AP} correlations of TTG rankings over all pairs of ad-hoc runs. It shows that there are strong correlation values for all of the evaluation metrics, especially recall and precision. There are some noticeable difference in the values of τ and τ_{AP}, where the latter is smaller. This is expected since τ_{AP} is more sensitive to changes on the ranks at the top of the list. According to Voorhees [8], a τ correlation over 0.9 "should be considered equivalent since it is not possible to be more precise than this. Correlations less than 0.8 generally reflect noticeable changes in ranking". A later study by Sanderson and Soborroff [7] showed that τ gets lower values when lists of smaller range of values are compared, which holds in our case. Thus, the correlation values achieved in Table 4 show that ranking of TTG systems is almost equivalent by all TTG evaluation scores regardless of the ad-hoc run used.

This finding is of high importance, since it suggests a possible solution to achieve less-biased evaluation of the TTG task, simply by using a common/standard ad-hoc run when evaluating new TTG systems.

One possible and straightforward ad-hoc retrieval run that can be used as a standard run for evaluating different TTG systems is the baseline run a_{13}. Such run is easy to construct through searching the tweets collection without any processing to the queries. Although the retrieval effectiveness of a_{13} is expected to be one of the poorest (see Table 1), when we calculated the average

τ and τ_{AP} correlation for ranking TTG systems with this run compared to the other 12 add-hoc runs using the wF_1 score, the values were 0.88 and 0.82 respectively. This is a high correlation according to Voorhees [8].

5 Discussion and Recommendation

In this study, we used a set of 13 ad-hoc retrieval runs and 8 TTG systems, resulting in a set of 104 different TTG outputs, which is a reasonable number for getting reliable results. Our main motivation behind the study was to investigate a potential bias of the currently-used TTG evaluation methodology, which is a critical and essential issue for future contributions to the task using the same dataset and evaluation methodology. The investigation confirmed the concern about the dependency of TTG output on the quality of the retrieval step. Nevertheless, we found that using one common ad-hoc retrieval run, fed to all TTG systems, might be sufficient for ranking these systems in a less-biased way using the current evaluation measures.

We recommend to use the baseline retrieval run (the one obtained using TREC search API) as the common run. It can be utilized in addition to other retrieval runs to allow for comparing TTG algorithms more fairly. Besides, using a high quality ad-hoc run continues to be highly recommended for understanding the performance of combining both retrieval and TTG methods for the best performing overall system pipeline.

Acknowledgments. This work was made possible by NPRP grant# NPRP 6-1377-1-257 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors (The first author was not funded by the grant.).

References

1. Buckley, C., Voorhees, E.M.: Evaluating evaluation measure stability. In: SIGIR (2000)
2. Fan, F., Qiang, R., Lv, C., Xin Zhao, W., Yang, J.: Tweet timeline generation via graph-based dynamic greedy clustering. In: AIRS (2015)
3. Hasanain, M., Elsayed, T.: QU at TREC-2014: Online clustering with temporal and topical expansion for tweet timeline generation. In: TREC (2014)
4. Hasanain, M., Elsayed, T., Magdy, W.: Improving tweet timeline generation by predicting optimal retrieval depth. In: AIRS (2015)
5. Lin, J., Efron, M., Wang, Y., Sherman, G.: Overview of the TREC-2014 microblog track. In: TREC (2014)
6. Magdy, W., Gao, W., Elganainy, T., Wei, Z.: QCRI at TREC 2014: Applying the KISS principle for the TTG task in the microblog track. In: TREC (2014)
7. Sanderson, M., Soboroff, I.: Problems with Kendall's tau. In: SIGIR (2007)
8. Voorhees, E.M.: Evaluation by highly relevant documents. In: SIGIR (2001)
9. Xu, T., McNamee, P., Oard, D.W.: HLTCOE at TREC 2014: Microblog and clinical decision support. In: TREC (2014)
10. Yilmaz, E., Aslam, J.A., Robertson, S.: A new rank correlation coefficient for information retrieval. In: SIGIR (2008)