TREC 2013 Temporal Summarization

Draft Metrics

April 22, 2013

1 Introduction

Each event will be retrospectively analyzed for important sub-events or ‘nuggets’, each with a precise timestamp and text describing the sub-event. Our evaluation metrics will measures the degree to which a system can generate these updates in a timely manner.

2 Notation

Formally, define the space of updates as,

$\mathcal{M}$	set of possible update messages
$\mathcal{T}$	time range of the evaluation
$\mathcal{U} = \mathcal{M} \times \mathcal{T}$	superset of updates

Given an event, our annotation process generates a set of gold standard updates or *nuggets*,

$\mathcal{N} \subset \mathcal{U}$	set of reference updates

We note the following,

- annotations are retrospective and subject to assessor error in the precision of timestamps
- an annotated message will be a very short sentence, including only a single sub-event

A system generates a set of timestamped updates generated in the manner described in the Guidelines,

$\mathcal{S} \subset \mathcal{U}$	set of system updates

3 Evaluation

Our goal in this evaluation is to measure the precision, recall, timeliness, and novelty of updates provided by a system.

3.1 Preliminaries

Our evaluation metrics are based on the following auxiliary functions. First, each message has an associated reward or utility,

$\mathbf{R} : \mathcal{M} \rightarrow \Re$	message reward function

This measures the importance of the sub-event in a summarization system. Second, each message has an associated cost,

$$\mathbf{C} : \mathcal{M} \to \Re \qquad \text{message cost function}$$

which can be used to introduce, for example, reading effort for longer updates. Finally, we also define a very important *matching function* between an update and an update set,

$$\mathbf{M}(u) = \text{argmin}_{\{u' \in \mathcal{S} : u_n \approx u'_n\}} u'_t \qquad \text{match time function}$$

which should be interpreted as 'given u, the earliest matching update in the $\mathcal{S}$'.

3.2 Online Metrics

We are interested in systems that provide updates that are relevant, comprehensive, novel, and timely. Our online metrics measure the performance of a system at a particular point in time during the simulation.

In order to measure *relevance* of system updates, we compute the fraction of updates which contain target nuggets. We refer to this as precision, defining it below,

$$\mathcal{P}(\tau) = \frac{\sum_{u \in \mathcal{N} : \mathbf{M}(u)_t < \tau} \mathbf{C}(u_n) \times \mathbf{R}(u_n)}{\sum_{u \in \mathcal{S} : u_t < \tau} \mathbf{C}(\mathbf{M}(u)_n)} \qquad \text{precision at time } \tau$$

We note that a system only gets credit for the earliest update matching a nugget; later updates matching that nugget will be treated as non-relevant. This captures the *novelty* objective and penalizes systems which return redundant updates.

In order to measure *comprehensiveness* of system updates, we compute the fraction nuggets which a system captures. We refer to this as recall, defining it below,

$$\mathcal{R}(\tau) = \frac{\sum_{u \in \mathcal{N} : \mathbf{M}(u)_t < \tau} \mathbf{R}(u_n)}{\sum_{u \in \mathcal{N}} \mathbf{R}(u_n)} \qquad \text{recall at time } \tau$$

If we consider the nugget timestamp as the earliest possible time at which we can detect a sub-event, then we also define a strict version of recall,

$$\mathcal{R}_{\text{s}}(\tau) = \frac{\sum_{u \in \mathcal{N} : (u_t < \tau) \wedge (\mathbf{M}(u)_t < \tau)} \mathbf{R}(u_n)}{\sum_{u \in \mathcal{N} : u_t < \tau} \mathbf{R}(u_n)} \qquad \text{strict recall at time } \tau$$

In order to measure *timeliness* by comparing the earliest matching updates to the nugget timestamps. The timeliness of an update is how long its timestamp is after (or before) the timestamp of the matching nugget. We then average the timeliness over matched nuggets,

$$\mathcal{T}(\tau) = \frac{\sum_{u \in \mathcal{N} : \mathbf{M}(u)_t < \tau} g(u_t - \mathbf{M}(u)_t)}{|\{u \in \mathcal{N} : \mathbf{M}(u)_t < \tau\}|} \qquad \text{timeliness at time } \tau$$

where g is a monotonically *increasing* function of $u_t - \mathbf{M}(u)_t$.