TREC 2018 News Track Overview

Ian Soboroff, Shudong Huang, Donna Harman
NIST

November 2018

Abstract

The News track is a new track for TREC 2019, focused on information retrieval in the service of helping people read the news. In cooperation with the Washington Post[1], we released a new collection of 600,000 news articles, and crafted two tasks related to how news is presented on the web.

1 Motivation

While news content has been a common genre in IR experimentation for a very long time, the evaluation tasks in IR have rarely if ever supported the "news user" – a consumer of news that is not an analyst. According to Pew Research studies, in 2016, roughly 38% of Americans got their news online, with the fraction increasing for younger consumers,[2] and in 2018 93% of American adults get at least some of their news online.[3] Pew further found in 2017 that at least two-thirds of Americans get news at least occasionally through social media.[4]

Moreover, since online delivery of news has shifted the focus away from the provider or publisher towards the story, news production has been dramatically democratized. If everyone can produce professional looking news, then understanding the context and background of information becomes a harder task for the consumer. In conjunction with The Washington Post, we are developing tasks around how news is presented on the web and thinking about how to enhance that learning experience.

[1] Certain companies and/or products are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the company or product identified are necessarily the best available for the purpose.

[2] `http://www.journalism.org/2016/07/07/the-modern-news-consumer/`

[3] `http://www.journalism.org/fact-sheet/digital-news/`, as of 6 Jun 2018

[4] `http://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/`

2 Data

The data for the News track is the TREC Washington Post Collection.[5] This collection contains five years of articles, from 2012 to 2017. The more than 600,000 documents in the collection comprise all Washington Post content: articles, columns, and blogs.

The documents are stored in "JSON-lines" format, that is, each document is a single long line of JSON. The articles are broken into content paragraphs, with interspersed media such as images and videos referenced by URL. Those URLs point back to the Washington Post website and according to the Post should persist at those URLs for the foreseeable future. This unique multimedia article format is novel for TREC but this track is not yet exploring it.

There are quite a few duplicate documents in the collection, because at times the Post will republish an article, and the provenance history is not represented in the data. We cleaned the collection to remove documents with identical content (including the document identifier). There are numerous other near-duplicates, and the track has not yet decided how to treat those articles.

The track shared topics with the Common Core track, as will be described more fully below with respect to each task. Half of the topics were re-used from older TREC collections, with a verification step taken to confirm that the topics had some relevant documents in the Post collection, but not too many (as was the case when topics were reused in the AQUAINT and NYT collections). The other half of the topics are newly developed on the Post collection. Taken together, the News and Common Core track topics and relevance judgments over all three tasks (adhoc, background linking, entity ranking) offer a novel assortment of training data for news systems.

With thanks due to Laura Deitz, we also provided a CAR-track formatted dump of Wikipedia articles as of August 2017, the end of the collection epoch. The Wikipedia dump was primarily for the entity ranking task but participants were free to use it however they liked.

3 Background Linking task

The goal of the background linking task is to develop evaluation data to support researchers in developing systems that can help users contextualize news articles as they are reading them. For example, news websites nearly always link to related articles in a sidebar, at the end of an article, from within the text of the article, or all three. We want to look at a particular case for linking: given that the user is reading a specific article (the query article), algorithms should recommend articles that this person should read next that are the most useful for providing context and background for the query article. The background article can be dated before or after the query article, because we are considering the use case where the user is reading the query article *now*, irrespective of when

[5] `https://trec.nist.gov/data/wapost/`

it was published, and the system is recommending background reading live at the time when the user is reading the query article.

It's important to note that links present in the Washington Post article collection are not training data for this task. In our conversations with the Post, their current practice is largely driven by the author of the article and does not follow any fixed guidelines or goal. Hence, we are designing this task as a specific kind of news recommendation task that would be useful in any news reading context, including the Post's website.

From our conversations with Post journalists about linking for background and context, every author has their own guidelines in their head, but three common rules emerged:

1. No wire service articles. (That is, from Associated Press (AP), AFP, etc)
2. No opinion or editorials.
3. The list of links should be diverse.

The corpus should not contain any wire service articles, so (1) is taken care of for free.[6] For (2), we decree that articles from the "Opinion", "Letters to the Editor", or "The Post's View" sections, as labeled in the "kicker" field, are not relevant. (3) is complicated as we are not sure we yet have a good understanding of diversity in the news recommendation context.

The topics for the Common Core track are being used as starters for the background linking task topics: query articles were selected based on documents found by the NIST assessors during Core track topic development.

```
<top>
<num> Number: 321 </num>
<docid>9171debc316e5e2782e0d2404ca7d09d</docid>
<url>https://www.washingtonpost.com/news/worldviews/wp/2016/
09/01/women-are-half-of-the-world-but-only-22-percent-of-its
-parliaments/</url>
</top>
```

The topic field "Docid" references the "id" field in the Washington Post corpus documents. "Url" references the "article_url" field in the documents. Both indicate the query article. This topic corresponds to the Core track topic 321, "Women in Parliaments", originally developed for TREC-6. For topics re-used from older collections, participants were permitted (in both the News and Common Core tracks) to make use of past relevance judgments and relevant documents from those collections.

The relevance scale used by the NIST assessors was:

0. The linked document provides little or no useful background information.

[6] There are actually some wire service articles, and we plan to cull them out in a future release of the collection.

1. The linked document provides some useful background or contextual information that would help the user understand the broader story context of the query article.

2. The document provides significantly useful background ...

3. The document provides essential useful background ...

4. The document MUST appear in the sidebar otherwise critical context is missing.

Systems retrieved up to 100 documents per topic and returned results in the standard trec_eval format. We pooled News and Core track runs together for assessment, to allow for the possibility of relevant background articles coming from Core track runs. We made an assumption that only documents relevant to the Core track topic could possibly qualify as background linking material, and hence assessments for the News track were only made on documents judged relevant (or not relevant but possibly having background on the topic) for Core.

This assumption turned out to be problematic: as we discovered after computing results, it was indeed possible for relevant background articles to be irrelevant to the Core track topic. For example, topic 809's title is "protect Earth from asteroids". The background linking query article for this topic is entitled "Europe will send a rover to Mars but won't protect Earth from an asteroid"; the EU decided to replace the destroyed ExoMars rover, but not to fund asteroid protection and hence this article is relevant to the Core track topic. Document `c6d6a0`[7], pooled from a News track run, is entitled "ESA confirms the ExoMars lander crashed, possibly exploded on impact"; this document was judged irrelevant to the Core track topic, and consequently was not judged for the News track, but is valuable background for the News track query article.

In reviewing the data, 2,155 pooled documents were found to have been judged not relevant to the Core topic, an average of 43 per topic. NIST staff rejudged those documents according to the News track background linking criteria above. While this undoubtedly introduced some inconsistency due to multiple assessors, we decided that this approach was the best of all available alternatives. The rejudging process added 250 level-1 relevant documents ("provides some useful background ..."), 66 level-2 relevant documents, 30 level-3, and 14 level-4. 1795 rejudged documents were judged as having little or no useful background information.

The primary metric for the background linking task is nDCG@5, with the gain value as 2^r where r is the relevance level from the scale above, and the zero relevance level contributing no gain. The evaluation reported all the standard trec_eval measures to a depth of 100. Figure 1 plots the nDCG@5 scores.

Five teams participated in the background linking task: Anserini (University of Waterloo); SINAI (Universidad da Jaen, Spain); UMass (University of Massachusetts, Amherst); htwsaar (Hochschule fuer Technik und Wirtschaft

[7]The full document ID is `c6d6a0695fcbc55330cc378d79301d0c`

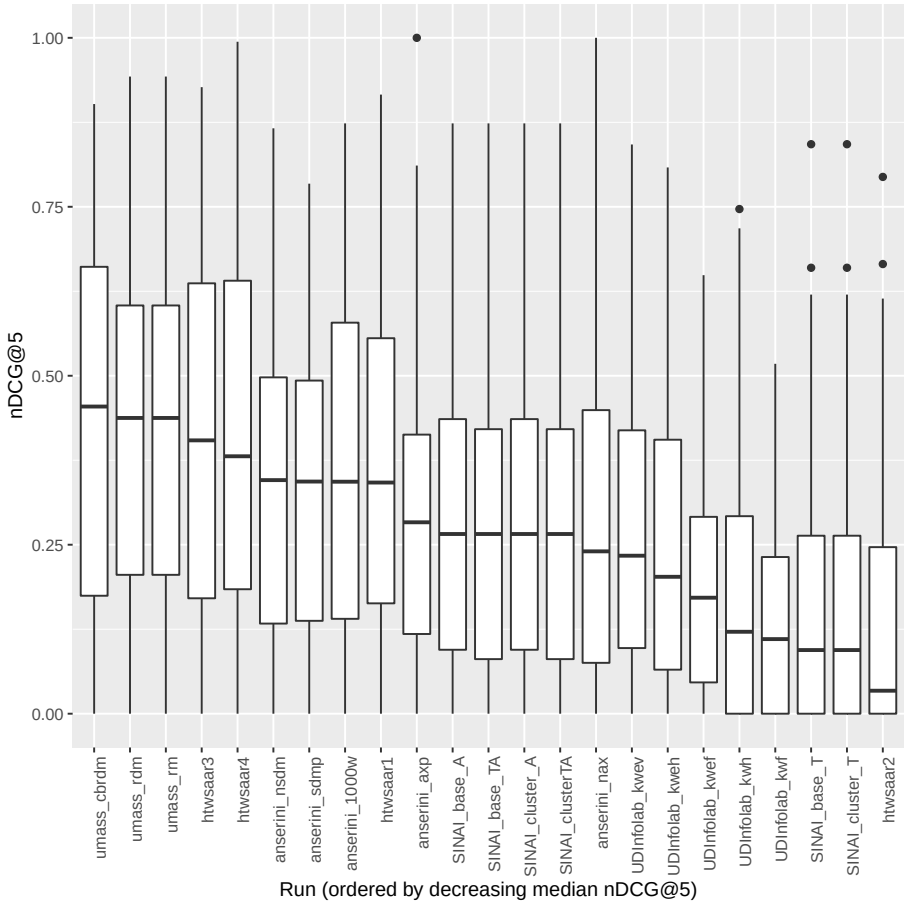

Figure 1: Boxplots for nDCG@5 score for each run in the background linking task. The plot illiustrates the median and interquartile distance across topics. Runs with overlapping boxes may not be statistically significantly different from one another.

des Saarlandes, Germany); and udel_fang (University of Delaware). They are described in Table 1.

4 Entity Ranking Task

In addition to providing links to articles that give the reader background or contextual information, journalists sometimes link mentions of concepts, artifacts, and entities to internal or external pages with in depth information that will help the reader better understand the article. For this second task, entity ranking, we automatically extracted named entities from query articles using Stanford's CoreNLP web service[8], and manually linked those entities to the provided Wikipedia dump. The task for systems was to rank the entities in order of importance – if providing a link to that entity would support the reader's understanding of the article or its broader context.

Part of the entity ranking task version of the "Women in Parliament" topic is as follows:

```
<top>
<num> Number: 321 </num>
<docid>9171debc316e5e2782e0d2404ca7d09d</docid>
<url>https://www.washingtonpost.com/news/worldviews/wp/2016/
09/01/women-are-half-of-the-world-but-only-22-percent-of-its
-parliaments/</url>
<entities>
  <entity>
  <id> 321.1 </id>
    <mention>Hassan Rouhani</mention>
    <link>enwiki:Hassan%20Rouhani</link>
  </entity>
  <entity>
  <id> 321.2 </id>
    <mention>Africa</mention>
    <link>enwiki:Africa</link>
  </entity>
  <entity>
  <id> 321.3 </id>
    <mention>Phumzile Mlambo Ngcuka</mention>
    <link>enwiki:Phumzile%20Mlambo-Ngcuka</link>
  </entity>
...
```

Note the "docid" and "url" sections are identical to those in the background linking task. Following those sections is a list of entities, linked to Wikipedia if the link is present in the dump. Every topic has at least three entities and one topic (414) has fifty-two entities.

[8] http://corenlp.run/

Systems returned a ranking of the entity IDs in the topic, using whatever resources they chose. The NIST assessors judged the entire set of entities for each topic on the following scale:

0. The linked entity provides little or no useful background information.
1. The linked entity provides some useful background or contextual information that would help the user understand the broader story context of the query article.
2. The entity link provides significantly useful background ...
3. The entity link provides essential useful background ...
4. The entity link MUST appear in the sidebar otherwise critical context is missing.

Two groups participated in the entity ranking task. Table 2 lists them, their properties, and their scores.

The primary metric for this task was the average precision of the entity ranking. In the "real world" task a system would need to cut off the ranking so as not to link unimportant entities, but in this first iteration of the task we did not measure selecting the cutoff point.

5 Conclusion

At the completion of year one of the News track, we have a first set of topics with judgments along a "news utility" scale in parallel with traditional TREC relevance judgments. Many questions remain: is the rating scale reasonable? Are the assessors distinguishing the levels meaningfully, or should the judgments be more categorical? Are these all "good topics" for these tasks, or are there some topics and article choices that are just inherently better for background linking? How can we characterize the set of good topics for these tasks?

Since it is the first year, we feel that the relevance judgments are likely to be a work in progress. It would be very interesting to see if systems tuned to these topics outperform other approaches that ignore these topics next year.

We plan to continue the track in TREC 2019, with fifty new topics. It is likely the track will incorporate the adhoc relevance judgment collection from the Common Core track. Although the News track will not feature an adhoc task with traditional TREC relevance criteria, "topical relevance" is still a useful concept for making news track assessments, and we remain interested in how judgments along these different criteria compare.

Group	Run	Type	Wiki?	Ext?
htwsaar	htwsaar1	auto	no	notused
htwsaar	htwsaar2	auto	no	used
htwsaar	htwsaar3	auto	no	used
htwsaar	htwsaar4	auto	no	notused
Anserini	anserini_1000w	auto	no	notused
Anserini	anserini_nsdm	auto	no	notused
Anserini	anserini_nax	auto	no	notused
Anserini	anserini_sdmp	auto	no	notused
Anserini	anserini_axp	auto	no	notused
udel_fang	UDInfolab_kweh	auto	no	used
udel_fang	UDInfolab_kwh	auto	no	used
udel_fang	UDInfolab_kwef	auto	no	used
udel_fang	UDInfolab_kwf	auto	no	used
udel_fang	UDInfolab_kwev	auto	no	used
UMass	umass_cbrdm	auto	no	notused
UMass	umass_rdm	auto	no	notused
UMass	umass_rm	auto	no	notused
SINAI	SINAI_base_A	auto	no	notused
SINAI	SINAI_base_T	auto	no	notused
SINAI	SINAI_base_TA	auto	no	notused
SINAI	SINAI_cluster_A	auto	no	notused
SINAI	SINAI_cluster_T	auto	no	notused
SINAI	SINAI_clusterTA	auto	no	notused

Table 1: Runs submitted to the background linking task. All runs were of type "auto" but could have been "manual" or "fdbk" indicating use of past TREC documents for older topics. "Wiki?" indicates if the run used the Wikipedia dump. "Ext?" indicates if the run used external resources.

Group	Run	Type	Wiki?	Ext?	MAP	nDCG@5
signal	signal-ucl-slst	auto	no	used	0.6894	0.5772
signal	signal-ucl-sel	auto	no	used	0.7158	0.6071
signal	signal-ucl-eff	auto	no	used	0.7144	**0.6084**
trema-unh	UNH-ParaBm25Ecm	auto	yes	notused	0.6828	0.3261
trema-unh	UNH-ParaBm25	auto	yes	notused	**0.7859**	0.4278
trema-unh	UNH-TitleBm25	auto	yes	notused	0.7741	0.4220

Table 2: Runs submitted to the entity ranking task. All runs were of type "auto" but could have been "manual" or "fdbk" indicating use of past TREC documents for older topics. "Wiki?" indicates if the run used the Wikipedia dump. "Ext?" indicates if the run used external resources. MAP and nDCG at rank 5 were the main metrics.