TREC 2019 News Track Overview

Ian Soboroff, Shudong Huang, Donna Harman
NIST

March 2020

Abstract

The News track focuses on information retrieval in the service of helping people read the news. In 2018, in cooperation with the Washington Post[1], we released a new collection of nearly 600,000 news articles, and crafted two tasks related to how news is presented on the web: background linking and entity ranking. This second iteration of the track continues these two tasks with minor updates.

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.

The initial run of the track in TREC 2018 introduced two tasks: recommending articles to read as background in the context of the article a user is

[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/`

presently reading (*background linking*) and ranking the entities mentioned in the article in order of how important it is to link to a Wikipedia page for that entity in order to provide background context (*entity ranking*). [2] This year continues these two tasks to build a combined set of 107 background linking topics and 102 entity ranking topics.

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 595,037 documents in the collection comprise all Washington Post content published in that interval: 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.

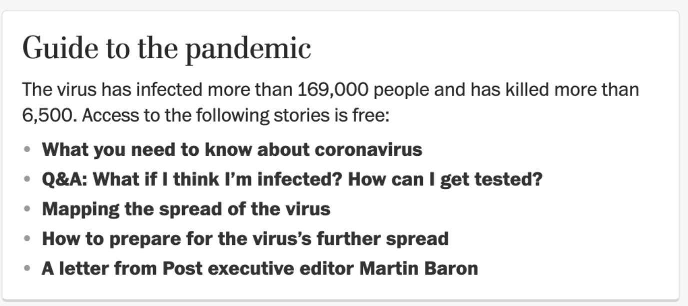
There are quite a few duplicate documents in the collection, because at times the Post will republish an article, and the provenance history is not represented in the data. We cleaned the collection to remove documents with identical content (including the document identifier). There are numerous other near-duplicates, and the overview this year considers their impact.

With thanks due to Laura Dietz, we provided a CAR-track formatted dump of Wikipedia articles as of August 2017, the end of the collection epoch. The Wikipedia dump was primarily intended 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. 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 it was published, and the system is recommending background reading live at the time when the user is reading the query article.

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



Guide to the pandemic

The virus has infected more than 169,000 people and has killed more than 6,500. Access to the following stories is free:

- **What you need to know about coronavirus**
- **Q&A: What if I think I'm infected? How can I get tested?**
- **Mapping the spread of the virus**
- **How to prepare for the virus's further spread**
- **A letter from Post executive editor Martin Baron**

Figure 1: A set of manually-curated background links from the Washington Post website (image taken on March 16, 2020).

Sometimes the Post will develop a link block in the context of a large story, for example the block shown in Figure 1 for a story during the Trump impeachment inquiry, a significant news event with many details and sub-events. This task imagines automatically providing links like these, for any article the user might be reading.

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 assessors following NIST's standard adhoc topic development process, which involves scouting the collection for a topic, loosely estimating the number of relevant articles that might exist, and crafting a title, description, and narrative statement. As a final step, the assessor selected a relevant article that they

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

had found during the development process as the query article for the topic. A total of sixty topics were developed and released for the evaluation.

```
<top>
<num> Number: 826 </num>
<docid>96ab542e-6a07-11e6-ba32-5a4bf5aad4fa</docid>
<url>https://www.washingtonpost.com/sports/nationals/the-minor-
leagues-life-in-pro-baseballs-shadowy-corner/2016/08/26/96ab542e-
6a07-11e6-ba32-5a4bf5aad4fa_story.html</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. In the first running of the track last year, the topics were shared with the Common Core track and some of those were re-used from previous TRECs, but this year all topics are new.

The relevance scale used by the NIST assessors was:

0. The linked document provides little or no useful background information.
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. The top 50 documents from each run were pooled for assessment. Three topics, 868, 877, and 881, were not fully judged in the allotted assessment period, so the evaluation is computed over 57 topics.

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. This is implemented for trec_eval by having the relevance levels in the qrels file be the gain values for nDCG. The evaluation reported all the standard trec_eval measures to a depth of 100. Figure 2 plots the nDCG@5 scores.

Nine teams participated in the background linking task:

- cityuni (City University of London)
- CLAC_NEWS_2019 (Computational Linguistics at Concordia)
- ICTNET (Institute of Computing Technology, Chinese Academy of Sciences)

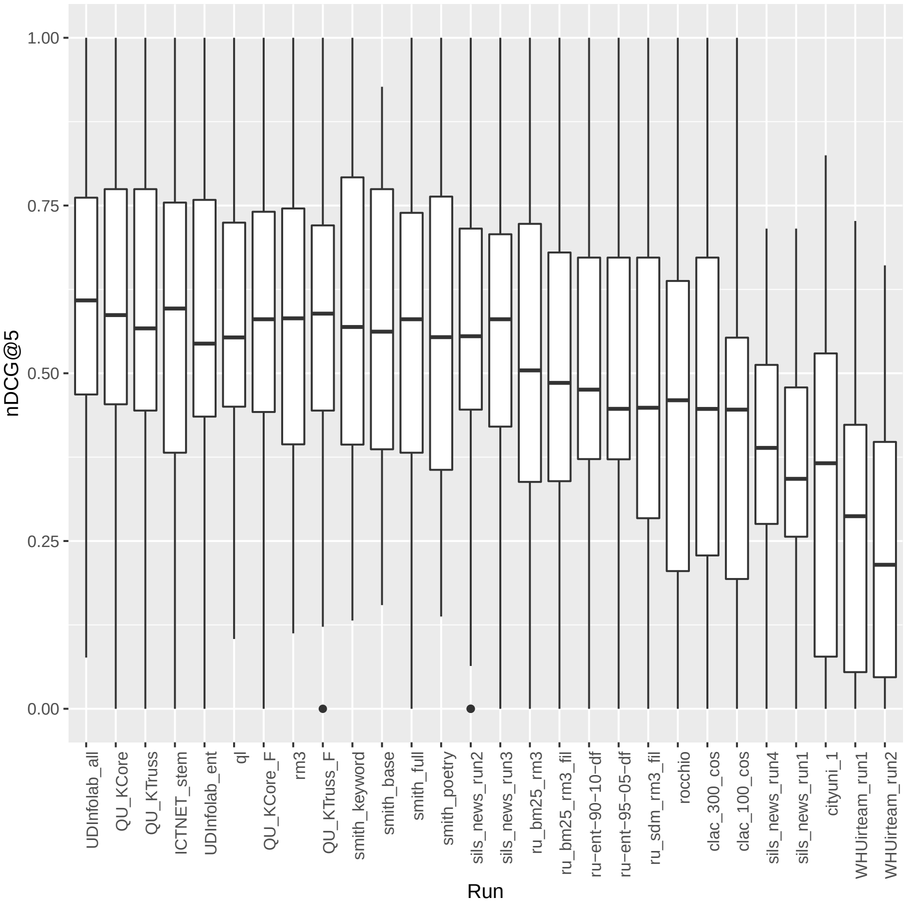


Figure 2: Boxplots for nDCG@5 score for each run in the background linking task. The plot illiustrates the median and interquartile distance across topics. Runs with overlapping boxes may not be statistically significantly different from one another.

- QU (Qatar University)
- QUARTZ_ITN (University of Padova)
- RUIR (Radboud University Data Science)
- Smith (Smith College)
- udel_fang (University of Delaware (InfoLab)).
- UNC_SILS (University of North Carolina at Chapel Hill)
- YQW2018CGroup (Big Data and Information Retrieval Group, Wuhan University)

28 runs were submitted for the background linking task.

3.1 Duplicate Documents

The Washington Post collection has many duplicate documents, covering a wide range of duplication from exact byte-match entries down to the document identifier up to copies and insertions. In 2018, NIST released an updated "v2" of the collection that removed exact duplicates, shrinking the collection from 608,180 documents to 595,037. The 2019 Conversational Assistance track published a list of near-duplicate paragraphs in the collection, as measured by Jaccard similarity of tokens.[7]

Duplicate documents can be problematic for evaluation: systems index multiple copies of the content, and may return duplicate results in the ranking, which can in turn lead to inconsistencies in relevance assessments if those documents are not cross-checked. In the case of near-duplicates, where the content may not exactly match due to edits, containment, copying, or other reasons, near-duplicate clusters need to be checked manually to ensure that the relevance judgment given is appropriate to all members of the cluster.

To investigate this phenomenon in the Post collection, we implemented document fingerprinting using minhashing, and fast similarity search using locality-sensitive hashing, following the work of Broder and colleagues as described by Leskovec et al. [1]. Our implementation uses overlapping nine-token n-grams as shingles, where tokens are taken from the "content" blocks of the documents only, separated by whitespace with nonword characters and HTML tags removed with simple regular expressions. We declare documents to be near-duplicates if their Jaccard similarity exceeds 0.84. We found that including metadata such as authors, titles, and kickers in the content hashing increased false positives, because of empty documents and very short documents in the collection. Manual inspection of true positives and true negatives showed that we found duplicate articles effectively, without mistakenly combining articles that were not duplicates but were very similar, like weekly lists of local events. Our set of 517,530 equivalence classes is posted on the TREC website.

[7] `http://boston.lti.cs.cmu.edu/Services/treccast19/duplicate_description.txt`, as of 10/29/2019.

We first resolved any relevance judgment conflicts within a near-duplicate class. We did this by manually reviewing all possible conflicts and resolving them as consistently as possible within the judgment of the assessor. In future years, we plan to sort documents in the pool according to near-duplicate class, so that near-duplicates are judged together in clumps. These resolved judgments were reported as “final” results to participants. The Kendall’s tau rank correlation of the system ranking before and after duplicate resolution was 0.95 for nDCG and 0.99 for MAP.

We then ran an experimental evaluation where we removed near-duplicates from both the relevance judgments and the runs. Near duplicates in an equivalence class were dropped except for a single “cluster representative”, and the runs were modified such that the first occurrence of any document in a cluster was replaced by the representative, and later occurrences of documents in that cluster were dropped. The tau correlation between the final official evaluation and this “deduplicated” evaluation was 0.96 for nDCG and 0.87 for MAP.

This is what would happen if all participants had our near-duplicate cluster list and removed duplicates after retrieval, but actual results from systems could be different if deduplication happens at indexing time since corpus statistics could change. In any event, the evaluation script needs to know the mapping to compute scores because we don’t know which member of a class a system will retrieve.

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 topic 826 from above is as follows:

```
<top>
<num> Number: 826 </num>
<docid>96ab542e-6a07-11e6-ba32-5a4bf5aad4fa</docid>
<url>https://www.washingtonpost.com/sports/nationals/the-minor-
leagues-life-in-pro-baseballs-shadowy-corner/2016/08/26/96ab542e-
6a07-11e6-ba32-5a4bf5aad4fa_story.html</url>
<entities>
  <entity>
  <id> 826.1 </id>
```

[8] http://corenlp.run/

```
    <mention>Richmond</mention>
    <link>enwiki:Richmond,%20Virginia</link>
  </entity>
  <entity>
  <id> 826.2 </id>
    <mention>Boston College</mention>
    <link>enwiki:Boston%20College</link>
  </entity>
  <entity>
  <id> 826.3 </id>
    <mention>Connecticut</mention>
    <link>enwiki:Connecticut</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. Topic 826 has 35. Topic 838 has the most with 38 entities. This is somewhat smaller than last year, where the largest topic had 54 entities mentioned as identified by Stanford CoreNLP.

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.

For eight of the 60 topics, no entities were judged to be useful, and so the evaluation is reported over 52 topics. Eight groups submitted 22 entity ranking runs:

- CMU (Carnegie Mellon University (Callan))
- ICTNET (Institute of Computing Technology, Chinese Academy of Sciences)
- OzUNLP (Ozyegin University)
- RUIR (Radboud University Data Science)

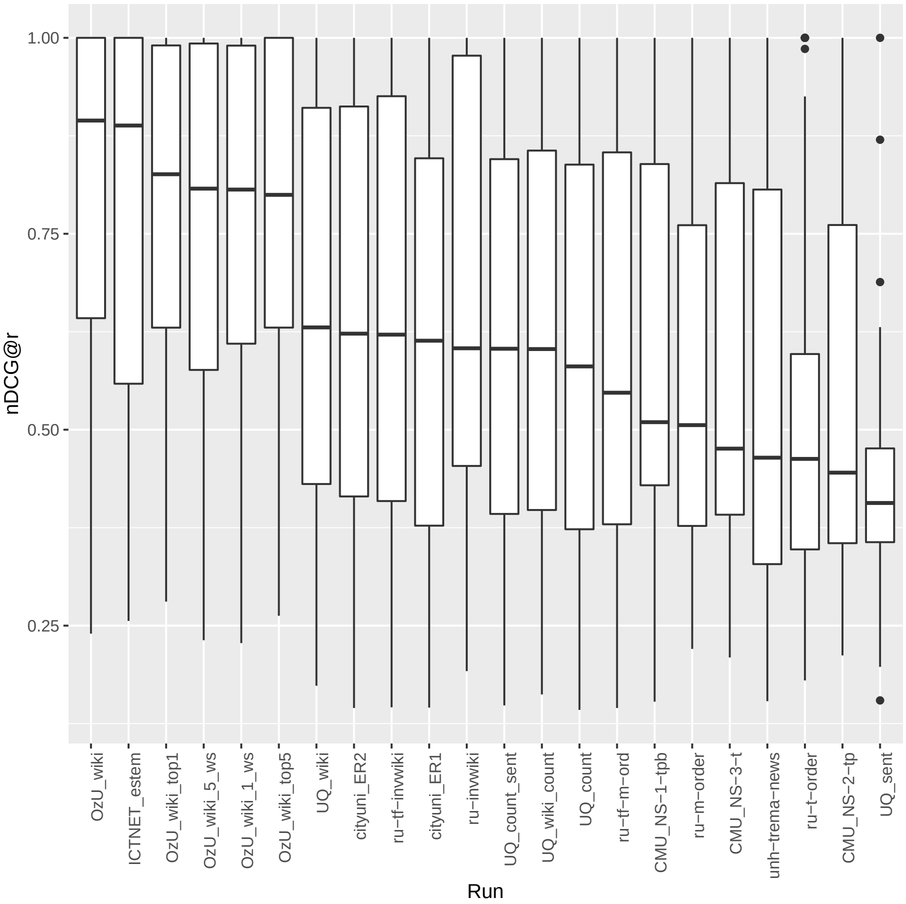


Figure 3: Boxplots for nDCG@r score for each run in the entity ranking task. The plot illiustrates the median and interquartile distance across topics. Runs with overlapping boxes may not be statistically significantly different from one another.

- TREMA-UNH (University of New Hampshire)
- UQ (University of Queensland)
- cityuni (City University of London)

Figure 3 shows scores for the runs ordered by average nDCG@r, where r is the number of relevant entities.

While last year's entity ranking evaluation reported average precision as the primary metric, this year we are reporting nDCG@r so as to take into account the different gain values for different entities. The Kendall's correlation $\tau = 0.91$ between nDCG@5 and nDCG@r, and $\tau = 0.88$ between nDCG@r and average precision. The lower correlation with MAP is an indication that there is an effect from the gain values on the scores.

5 Conclusion

The second year of the News track continued the background linking and entity ranking tasks to complete a combined set of more than 100 topics for these tasks. Last year we felt that the relevance assessment scale was somewhat experimental, but at this point the scale feels appropriate for the task and the assessors understand it.

Unfortunately, this year we were not able to include an adhoc task. The assessors composed adhoc topics, but we felt it was too confusing to judge according to both the adhoc and background linking criteria at the same time, and we did not have resources for two assessment passes over the pools. Thus whether the background linking task requires something beyond good adhoc ranking is an outstanding question for research.

The track will continue in TREC 2020, and plans for the track will be discussed in the planning workshop.

References

[1] Jure Leskovec, Anand Rajaraman, and Jeffrey Ullman. *Mining of Massive Datasets*. Cambridge University Press, 2nd edition, 2014. Downloadable at `http://www.mmds.org/`.

[2] Ian Soboroff, Shudong Huang, and Donna Harman. TREC 2018 news track overview. In *Proceedings of the 27th Text REtrieval Conference (TREC 2018)*, Gaithersburg, MD, November 2018.