

TREC 2020 News Track Overview

Ian Soboroff, Shudong Huang, Donna Harman
NIST

November 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 Post¹, 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. For 2020, we added more documents to the collection and retired the entity ranking task in favor of a new wikification task.

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 2018, 34% of U.S. adults said they preferred to get news online² In 2019, they found that about 20% of U.S. adults say they get political news primarily through social media³. They also found that readers getting their information on major news stories such as the coronavirus outbreak from social media tended to be less knowledgeable about those topics than those who received their news from other sources.

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 support the reader in understanding the news.

¹Certain 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<https://www.pewresearch.org/fact-tank/2019/09/11/key-findings-about-the-online-news-landscape-in-america/>

³<https://www.journalism.org/2020/07/30/americans-who-mainly-get-their-news-on-social-media-are-less-engaged-less-knowledgeable/>

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 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*). [1, 2] For the 2020 track, the entity ranking task was replaced with a new *wikification* task.

2 Data

The data for the News track is the TREC Washington Post Collection.⁴ This collection contains seven years of articles, from 2012 through 2019. The 671,947 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.

Based on findings in last year’s track, the collection has been cleaned of exact- and near-duplicate articles. Near-duplicate documents were detected using locality-sensitive hashing with minhash fingerprints. [2]

With thanks due to Laura Dietz, we provided a CAR-track formatted dump of Wikipedia articles as of the beginning of 2020, the end of the collection epoch. The Wikipedia dump was primarily intended for the wikification 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.

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 appearing during the

⁴<https://trec.nist.gov/data/wapost/>

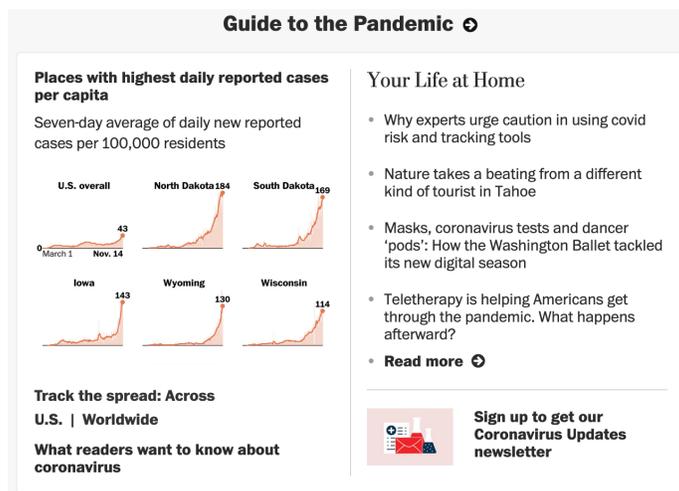


Figure 1: A set of manually-curated background links from the Washington Post website (image taken on November 14, 2020).

COVID-19 pandemic, 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.⁵ 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.

⁵There are actually some wire service articles, and we plan to cull them out in a future release of the collection.

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 had found during the development process as the query article for the topic. A total of fifty topics were selected by NIST and released for the evaluation.

```
<top>
  <num> Number: 890 </num>
  <docid>QD7AORAY6AI6TCH67H3XU06LNQ</docid>
  <url>https://www.washingtonpost.com/entertainment/music/as-a-
  classical-music-critic-i-used-to-think-the-star-wars-score-was-
  beneath-me-i-was-wrong/2019/01/17/80fe0744-18f0-11e9-88fe-
  f9f77a3bcb6c_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; this constitutes a *run*. The top 50 documents from each run were pooled for assessment. No relevant documents were found for topic 917, and thus it was dropped from the topic set.

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 `qrrels` 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:

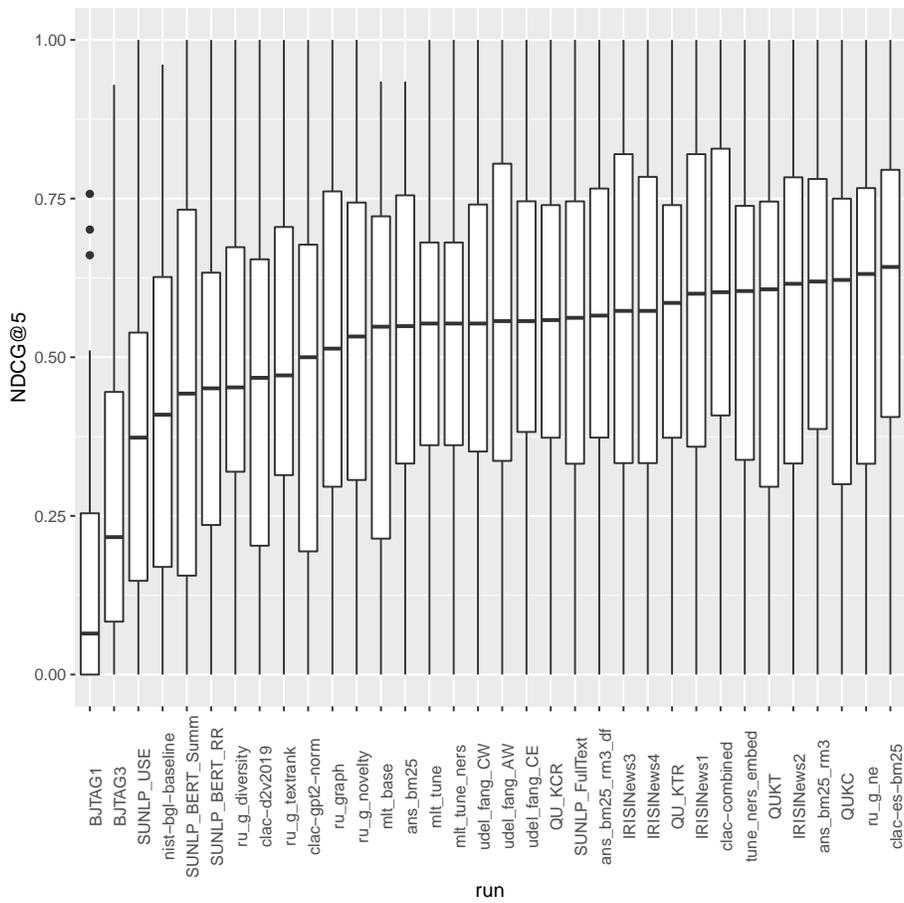


Figure 2: Boxplots for nDCG@5 score for each run in the background linking task. The plot illustrates the median and interquartile distance across topics. Runs with overlapping boxes may not be statistically significantly different from one another.

- BJTAG (SAS Research and Development (Beijing) Co., Ltd.)
- CLAC_NEWS_2020 (CLaC Lab (Computational Linguistics at Concordia))
- IRLABISI (Indian Statistical Institute)
- OSC (OpenSource Connections)
- QU (Qatar University)
- RUIR (Radboud University IR Research Group)
- SUNLP (Sabanci University)
- anserini (Anserini Gaggle)
- udel_fang (InfoLab at University of Delaware)

33 runs were submitted for the background linking task.

4 Wikification 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. In the first two years of the track, we worked on ranking entities within the document as to their priority for linking. For this year, the task was grown into a full *wikification* task.

Wikification is providing Wikipedia-like markup in a document by linking related content to contextually-relevant anchors. The goal of wikification depends on the application; in Wikipedia, the goal is to link articles to other articles as cross-references. For news reading, we imagine that wikification could be used to recommend interstitial links to the author or editor of the article, where those links serve a similar function to results in the background linking task: provide essential background and context to the reader of the article.

This definition of wikification differs in some important ways from the natural language processing field’s version, where it essentially means linking entities in the article (which may be a Wikipedia article) to other Wikipedia articles. First, we do not restrict the anchor to entity mentions but rather they can be any contiguous span of text within a block. Second, the target of the link can be the knowledge base (Wikipedia) or another article in the Post. Lastly, the task is formulated as ranked recommendation, so high performance is modeled by selecting useful anchors and linking them in a way that provides context to the user, instead of complete and correct entity linking.

The task is operationalized as follows: Given the same starting article as in the background linking task, systems were to provide a ranked list of link anchors and suggestions. A link anchor was a triple of content block number, starting

character, and anchor length, defining an extent of characters in the document. The suggested target for the anchor could be either another Washington Post article, or a Wikipedia article in the provided Wikipedia dump. Individual (anchor, target) pairs are ranked by the system-provided score as in a traditional TREC run.

At the time that the task was released to participants, the rubric for relevance judgments was not yet set, although the coordinators had expected the scheme to mirror that of background linking. However, as the assessment plan was developed, it was clear that a wikification result could be poor for multiple reasons. Thus, the following two-level, four-part scheme was used:

1. is the anchor **sensible**, that is, does it break on word boundaries and seem reasonable in a “lexical” sense? Anchors that break in the middle of words or awkwardly in the middle of phrases are not sensible.
2. is the anchor **useful**, that is, would a link at that place in the document possible help the reader better understand the article? Converting a noun phrase that has no bearing on the main topic of the article would be sensible, but not useful.
3. does the target **match** the anchor? Is the target document what the reader would reasonably expect to be linked from that anchor? Linking the text “Justin Bieber” to the Wikipedia page for Barack Obama would not match, although the anchor seems sensible and is conceivably useful.
4. is the target **helpful**? Does the content in the linked article provide background information or context for understanding the topic document? This is essentially the relevance criterion in the background linking task.

Three groups submitted eight runs to the wikification task which, given that at submission time the relevance and evaluation criteria were not well-defined, signifies true bravery and sacrifice in the name of research, and we salute them here (with appropriate disclaimers):

- BJTAG (SAS Research and Development (Beijing) Co., Ltd.)
- SUNLP (Sabanci University)
- TUW-IFS (Vienna University of Technology)

The top-ranked 50 (anchor, target) pairs were pooled for assessment.

Figure 3 shows scores for the runs ordered by average nDCG@5. Gain values were derived from the assessments as follows:

1. if the anchor is sensible and the target matches, the pair received a gain of 1.
2. if additionally the anchor is useful, the pair received a gain of 5.
3. if additionally the target is helpful, the pair received a gain of 10.

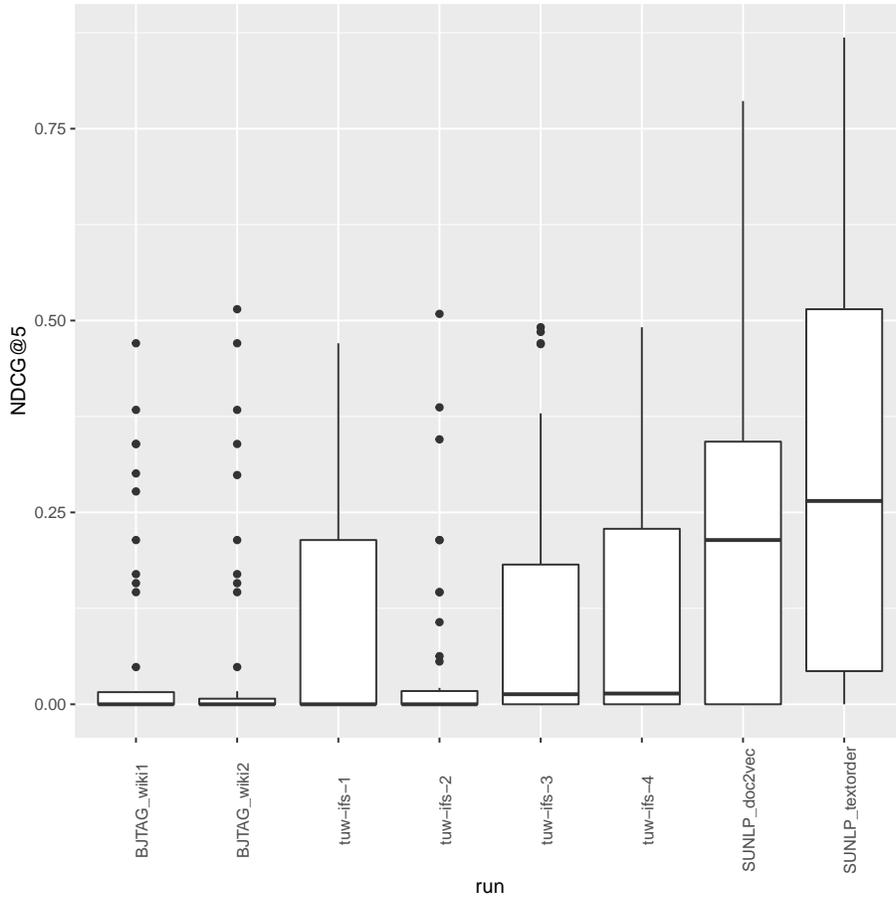


Figure 3: Boxplots for nDCG@5 score for each run in the wikification task. The plot illustrates the median and interquartile distance across topics. Runs with overlapping boxes may not be statistically significantly different from one another.

5 Conclusion

The third year of the News track continued the background linking task and introduced a new wikification task to expand the track notion of recommendation to interstitial links. Currently, the track is slated to continue in TREC 2021.

References

- [1] 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. <https://trec.nist.gov/pubs/trec27/papers/Overview-News.pdf>.
- [2] Ian Soboroff, Shudong Huang, and Donna Harman. TREC 2019 news track overview. In *Proceedings of the 28th Text REtrieval Conference (TREC 2019)*, Gaithersburg, MD, November 2019. <https://trec.nist.gov/pubs/trec28/papers/OVERVIEW.N.pdf>.