TREC 2021 News Track

Guidelines v1.1, 25 May 2021

Track coordinators: Ian Soboroff, Shudong Huang, Donna Harman (NIST)
Google group: https://groups.google.com/forum/#!forum/trec-news-track
Slack: trectalk.slack.com#news-2021 (restricted to active participants)
Register to participate at https://trec.nist.gov/pubs/call2021.html

Changelog

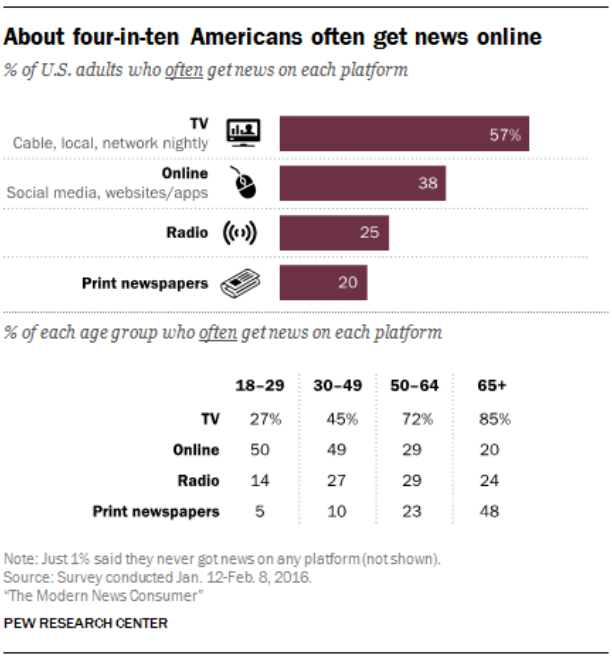
Here are the big changes for 2021:

- We have a new version of the Washington Post collection: v4 includes articles up through the end of 2020.
- The background linking task has a new faceted background linking subtask.

25-May-2021: Updated to include link to Wikipedia dump (same as 2020).

Motivation

While news and newswire 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 a Pew Research study in 2016, roughly 38% of Americans get their news online, with the fraction increasing for younger consumers.



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.

This track is not about detecting "fake news". Rather, we are envisioning new information access tools that help the user understand the context of a story, wherever they are reading it. 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 larger question is, what roles can IR play in this new, noisy, adversarial online news domain?

Data

The data for the track is the TREC Washington Post Collection, first released in 2018, and updated each year with new articles. The collection spans nine years of articles, 2012 - 2020, and you can get it at no cost from NIST after completing the requisite usage agreement. The data URL is https://trec.nist.gov/data/wapost/

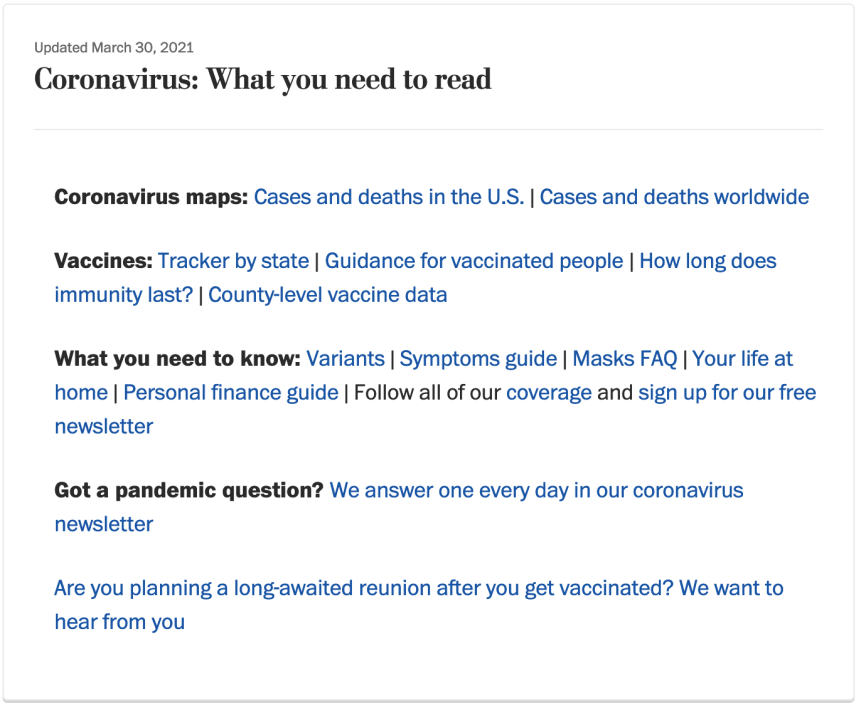
The files are "JSON-lines" format, that is, each document is a single long line of JSON. The articles are broken into content paragraphs, with interspersed media such as images and videos referenced by URL (for articles through 2019). Those URLs point back to the Washington Post website and should persist at those URLs for the foreseeable future.

For the wikification task, the Wikipedia dump we are using is available at http://trec-car.cs.unh.edu/datareleases/v2.4/car-wiki2020-01-01.tar.xz

Task 1: Background Linking

The goal background linking task is to develop 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), recommend articles that this person should read next that are the most useful for providing **context and background** for the query article.

The Post currently implements their take on this task, manually, using what I call "explainer boxes". Here is an example:



Updated March 30, 2021

Coronavirus: What you need to read

Coronavirus maps: Cases and deaths in the U.S. | Cases and deaths worldwide

Vaccines: Tracker by state | Guidance for vaccinated people | How long does immunity last? | County-level vaccine data

What you need to know: Variants | Symptoms guide | Masks FAQ | Your life at home | Personal finance guide | Follow all of our coverage and sign up for our free newsletter

Got a pandemic question? We answer one every day in our coronavirus newsletter

Are you planning a long-awaited reunion after you get vaccinated? We want to hear from you

This explainer box appeared at the end of most articles about COVID at the time that I wrote v1.0 of these guidelines. As you can see, it has something like a title/description, “Coronavirus: what you need to read”, and a set of linked articles grouped into categories. These categories reflect different reasons that the user might want to follow up and read more on this topic.

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 assessors will judge wire service articles as **not relevant**. (There are not a lot of these in the collection, but they’re hard to automatically remove.) 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**. For (3), we have a new condition for this task that includes “subtopics” that represent reasons for seeking background.

Results will be pooled and judged by NIST assessors on the following scale:

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 an explainer box or list of context links, otherwise critical context is missing.

Input

There will be 50 new topics for 2021. The topics will have the following format format:

```
<top>
<num>Number: xxx </num>
<docid>f30b7db4-cc51-11e6-a747-d03044780a02</docid>
<url>https://www.washingtonpost.com/local/public-safety/homicides-remain-steady-in-the-washington-region/2016/12/31/f30b7db4-cc51-11e6-a747-d03044780a02_story.html</url>
<title>Topic title</title>
<desc>I would like to learn more about this topic</desc>
<narr>
A traditional TREC narrative paragraph on the topic
</narr>
<subtopics>
  <sub num=”1”>This is the first subtopic.</sub>
  <sub num=”2”>And this is the second one.</sub>
</subtopics>
</top>
```

"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. The title/description/narrative sections describe the topic for purposes of the explainer box, and the subtopics are the subheadings of the box.

Systems may use any of the top-level fields (source article, title, description, and narrative), and additionally may be “subtopic” run or not. A “subtopic” run may use the subtopic descriptions and has a slightly different output format.

Output

Submissions should be standard TREC format, that is, trec_eval results file format:

```
1 Q0 2707e25a-cfaf-11e6-a87f-b917067331bb 1 37.5 myrun
1 Q0 513673ee-d003-11e6-b8a2-8c2a61b0436f 2 33.2 myrun
...
1 Q0 f8ded480-cdef-11e6-b8a2-8c2a61b0436f 99 0.5 myrun
2 Q0 350e3d74-cf94-11e6-a87f-b917067331bb 1 55.2 myrun
...
```

Systems may retrieve up to 100 documents per topic. The first field is the topic id ("<num>" in the topic), the second field is a literal "Q0", the third field is the document ID of the linked

document, the fourth field is the rank (ignored), the fifth field is the score, and the sixth field is the runtag. Note that trec_eval sorts by descending score and breaks ties using document IDs.

For “subtopics” runs, the topic field should be XXX.y, where ‘XXX’ is the topic number, and ‘y’ is the subtopic number.

We plan to pool documents from all runs together for assessment and subtopic assignment.

Metrics

The primary metric for the background linking task will be nDCG@10, where the gain value is 2^(*r*-1) where *r* is the relevance level from the scale above, and the zero relevance level contributes no gain. Evaluation will use trec_eval so all traditional TREC measures will also be reported to a measurement depth of 100.

Additionally, for subtopic runs we will report alpha-nDCG@10, and possibly other diversity metrics as capabilities allow.

Example

Query article: Love in the time of climate change: Grizzlies and polar bears are now mating (May 23, 2016)

This article describes and analyzes a phenomenon where grizzlies and polar bears are mating to create a new species known as pizzlies or grolars. It explains why this is happening and points out that it happens (or has happened) to other species as well. Articles along these lines are good background links. For example:

- Coywolves, coyote-wolf hybrids, are prowling Rock Creek Park and D.C. suburbs (July 1, 2014)
- Humans and Neanderthals may have interbred 50,000 years earlier than previously thought (February 17, 2016)

However, the following article is of less relevance and should be ranked lower because it’s not about interbreeding.

- Why do seals keep trying to have sex with penguins? (November 18, 2014)

Task 2: Wikification

In addition to providing links to articles that give the reader background or contextual information, journalists sometimes link mentions of concepts, artifacts, entities etc to internal or external pages with in depth information that will help the reader better understand the article.

Wikification is the automatic hyperlinking of entities, concepts, or references to another resource that provides more information on the linked thing, in the spirit of linking in Wikipedia. In Post articles, authors frequently allude to other Post articles, news articles from other sources, or external entities, which they link directly into the prose text of their articles. This task will explore whether systems can automatically wikify the Post source article usefully for the news reader, who as in Task 1 we envision as seeking background or contextual information on what they are reading. While for background linking, the context is the entire article, for wikification, the context is a short, explicit extent in the article that will be decorated with a hyperlink.

The input for this task will be the same topic statement as in the Background Linking task. There are two possible topic conditions. The first is for runs that only use the target article from the topic. The second is for runs that make use of any other fields in the topic.

Systems will provide output in the following line-oriented format:
`tid score cnum start len link`

Where

- `Tid` is the topic ID,
- `Score` is a numeric score for ranking,
- `Cnum` is the sequence number of the content block, starting from zero, of the linked passage,
- `Start` is the character offset in the content block, starting from zero, for the anchor text start,
- `Len` is the length of the anchor text, in characters. Len must be greater than zero.
- `Link` is the target of the hyperlink being proposed by the system.

Systems may link either to Washington Post articles in the collection, or to Wikipedia articles in the Wikipedia dump provided. A Post article link will be in the format `post_v4:docid`, and the semantics of a Post article link is that the anchor text is referring to or alluding to something covered or explained in the linked article, which would be useful to the reader in understanding or contextualizing the source article. A Wikipedia article link will be in the format `enwiki:name`, **following the identifiers in the Wikipedia dump**. The semantics of a Wikipedia link are as in the original entity ranking task: the linked article provides useful background or context that would benefit the reader in understanding the source article.

Ranked links will be pooled and judged by NIST assessors as follows:

- **Is the link anchor sensible?** Does the anchor highlight a sensible word or phrase? Does it make sense for the article to have a link in the text here, that would provide background or context in reference to the text in the anchor?
  For example, if the anchor isn't visible, or starts or ends in the middle or a word, or seems to cover an arbitrary part of a sentence, it's not sensible.
- **Is the link anchor useful?** Would a link there possibly help you better understand the topic document?
  For example, if the anchor text highlights an important concept or entity in the article, it might be useful to have a link there.
- **Does the target match the anchor?** Is the target document what you would expect to be linked from the anchor text?
  For example, if the anchor text is "Washington" but the target article is "New York", that doesn't match.
- **Is the target helpful?** Does the content in the linked article provide background information or context for understanding the topic document?
  For example, if the target article in fact provides useful context or background information for understanding the article.

Metrics

The primary metric for this task will be nDCG@10, where the gain value is 2^(*r*-1) where *r* is the relevance level from the scale above, and the zero relevance level contributes no gain. Evaluation will use trec_eval so all traditional TREC measures will also be reported to a measurement depth of 100.

Rules

Manual or automatic runs

When you submit your runs, you will be asked to indicate if the run is manual or automatic. An **automatic** run involves no manual intervention, but runs fully automatically from the topic file. In contrast, **manual** runs can involve human intervention – including manual query formulation, manual relevance feedback, and reweighting/reranking by hand.

Duplicate document handling

We have cleaned the document collection to remove exact and near-duplicate documents. Following adhoc practice, we will judge any remaining duplicate documents independently.

External resources

We are using the same Wikipedia dump as was used in 2020. This dump is formatted identically to dumps used in the CAR track. Wikipedia links in Task 2 will be with respect to this collection.

Past relevance judgments

Last year's topics and their assessments are available from the TREC data page (https://trec.nist.gov/data/). You can use them in tuning your systems.

Key dates

Guidelines released: April 2021

Topics released: April 30, 2021
Runs due: June 4th, 2021
TREC: November 15-19, 2021