

TREC 2019 News Track

Guidelines v1.2, 5 July 2019 (key dates updated)

Track coordinators: Shudong Huang, Ian Soboroff, Donna Harman (NIST)

Google group: <https://groups.google.com/forum/#!forum/trec-news-track>

Slack: [#news-2019](https://trectalk.slack.com) (restricted to active participants)

Register to participate at <https://trec.nist.gov/pubs/call2019.html>

Changelog

Here are the big changes for 2019:

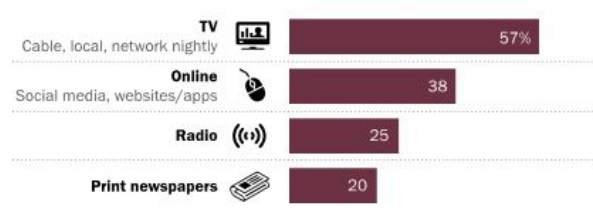
- TREC is using Slack for communication among active participants. You should make an effort to be on [#news-2019](https://trectalk.slack.com).
- There will be 50 new topics for this year, developed for the track. The topics and relevance judgments from last year are available for tuning.
- The tasks are the same: background linking and entity ranking.
- We are not making efforts to dedupe the collection beyond the removal of exact duplicates in v2 of the corpus.

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.

About four-in-ten Americans often get news online

% of U.S. adults who *often* get news on each platform



% of each age group who *often* get news on each platform

	18-29	30-49	50-64	65+
TV	27%	45%	72%	85%
Online	50	49	29	20
Radio	14	27	29	24
Print newspapers	5	10	23	48

Note: Just 1% said they never got news on any platform (not shown).
Source: Survey conducted Jan. 12-Feb. 8, 2016.

"The Modern News Consumer"

PEW RESEARCH CENTER

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, released last year. This is five years of articles, 2012 - 2017, 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. Those URLs point back to the Washington Post website and should persist at those URLs for the foreseeable future.

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 have cleaned all "exact-match" duplicates from the collection, but left near-duplicates for the moment.

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.

Note that links 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 assessors will judge wire service articles as **not relevant**. 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 have a good understanding of diversity in the news recommendation context; resources permitting, we will

pilot some annotations with the assessors which we hope will inform next year's version of the task. (We wanted to do this in 2018 but did not have the resources.)

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 the sidebar otherwise critical context is missing.

Input

There will be 50 new topics for 2019, developed on the Washington Post collection. The topics will mimic the standard TREC topic 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>
</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.

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.

Metrics

The primary metric for the background linking task will be $nDCG@5$, 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.

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: Entity Ranking

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. For this second task, entity ranking, we'll provide a wikipedia dump, extract all entities from each query article that have entries in the wikipedia and ask the system to determine which entities are linkable for the given article and rank them in terms of "usefulness". We will limit entity types to person, organization, location, geopolitical entity, and facility with proper names.

Consider [this Washington post news article](#). Stanford's CoreNLP web service (<http://corenlp.run/>) finds these named entities, among others:

- San Antonio

- Walmart
- James Matthew Brady Jr.
- Justice Department
- Mexicans
- Los Zetas
- The Washington Post

A system that could automatically link "James Matthew Brady Jr." to a Wikipedia page about the event, San Antonio to a map location, and Los Zetas to a background document would be providing valuable further reading to the user. In contrast, linking to Walmart's Wikipedia page or homepage would be less useful, and there is no reason to link the mention of the Washington Post in this article. As with determinations of relevance, reasonable people might disagree about which entities are best to link, making this a good task for ranked retrieval systems.

(Note that there are also many more entities which are not so simple for systems to identify, such as "A federal grand jury", "immigration authorities", "federal prison", and "as many as 200 people". For now, we are keeping things simple, but the end goal is automatic, importance-driven wikification of the article.)

There are three components to this task: identifying the entities to link, selecting the most important entities, and determining what to link them to. The first and last tasks, called Entity Detection and Linking (EDL) in the Text Analysis Conference, are the subjects of active research in the NLP community. For this task, we will focus on stage two: separating important entities from unimportant ones, for the purpose of making informational linkages for readers of the article.

Ranked entities will be pooled and judged by NIST assessors 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.

Input

Topics will be in a modified TREC format, as follows:

```
<top>
<num> Number: xxx </num>
<docno> 4989ebfeb752e6b317d1ef3997b21a01 </docno>
<url>https://www.washingtonpost.com/news/post-nation/wp/2017/08/17/official
s-trucker-indicted-could-face-death-penalty-after-10-migrants-die-in-smuggl
```

```

ing-incident/
```

```

</url>
<entities>
  <entity>
    <id> xxx.1 </id>
    <mention>San Antonio</mention>
    <mention>Alamo city</mention>
    <link>enwiki:link-into-wikipedia-dump</link>
  </entity>
  ...
</entities>
</top>

```

The "entities" block is a sequence of one or more entities, each with an ID, the mention string, its location in the document, and a link into a Wikipedia dump provided for the track.

Output

Systems will provide a ranking of the entities in the topic, in trec_eval format:

```

xxx q0 xxx.1 1 37.5 runtag
...

```

where the "docno" field should have an entity ID from the topic.

Metrics

The primary metric for this task will be nDCG@5, 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 duplicate documents, that is, duplicates down to the docid. However, we have not cleaned near duplicates. Following adhoc practice, we will judge duplicate documents independently.

External resources

We will provide a Wikipedia dump from August 20, 2017, which coincides with the end of the epoch of the Washington Post collection. This dump will be formatted identically to dumps used in the CAR track. Entity links in Task 2 will be with respect to this collection. A checkbox on the run submission form will ask if you made use of this dump.

Past relevance judgments

Last year's topics and their assessments are available from the TREC data page (<https://trec.nist.gov/data/news2018.html>). You can use them in tuning your systems.

Additional Examples

Topic 321: Women in Parliaments

Query article: [Another way Britain's vote made history: More women than ever before were just elected](#) (June 9, 2017)

Background links:

- [British prime minister calls for elections in June amid Brexit fallout](#) (April 18, 2017)

This is WaPo's initial report on May's call for a June snap general election. The query article is focused on record number of female MPs elected but doesn't provide much information on the election itself. So this article provides significant background to the reader.

- [The moving historical drama 'Suffragette' rings all too true today](#) (October 29, 2015)

The name "Emily Wilding Davison" has a prominent role in the query article. As it happens, there was a movie made in 2015 about the women's rights movement in the early 20th century and the movie's climax is based on the real life story of Emily Wilding Davison.

Entities:

- Theresa May
- Emily Wilding Davison
- U.K. Parliament
- U.K.
- Jessica Howard*
- Owen Barder*
- Preet Kaur Gill

- House of Commons
- Labour Party
- Birmingham Edgbaston
- Tory
- Delphine Desbenoit*
- London

*Note that these entities do not have wikipedia entries and thus will not be included in the list to be ranked by systems.

The most important entity for this article is undoubtedly “Emily Wilding Davison”. A reader might also be interested in knowing more about the UK political system, and so “House of Commons”, “Labour Party”, and “Tory” should also be high. Linking “UK” or “Theresa May” is probably not very informative for most readers.

Topic 809: Protect Earth from Asteroids

Query article: [Europe will send a rover to Mars but won't protect Earth from an asteroid](#)
(December 5, 2016)

This article is about the European Space Agency's success of securing funds to send a rover to Mars and its failure to secure funding to survey a near-Earth asteroid in order to understand how to deflect an asteroid coming towards the Earth.

Background links:

- [ESA confirms the ExoMars lander crashed, possibly exploded on impact](#)

The ExoMars lander's crash landing was a setback for ESA's space programs.

- [We really need to figure out how to stop a killer asteroid, scientists say](#)
- [NASA's mission improbable](#)

Entities:

- Europe
- Mars/Red Planet
- Earth
- European Space Agency/ESA
- Schiaparelli
- NASA
- Jan Woerner*
- Bruce Willis
- Reuters
- Lucerne
- Switzerland
- Don McCoy*
- Russian Federal Space Agency

- Trace Gas Orbiter
- Roberto Battiston*
- Italian Space Agency
- BBC

*Note that these entities do not have wikipedia entries and thus will not be included in the list to be ranked by systems.

Key dates

Guidelines released: May 2019

Background linking topics released: June 6, 2019

Background linking runs due: August 18, 2019

Entity ranking topics released: tbd

Entity ranking runs due: September 22, 2019

TREC: November 13 - 15, 2019