TREC 2018 News Track

Shudong Huang
100 Bureau Drive
Gaithersburg
Maryland 20899-8940
shudong.huang@nist.gov

Ian Soboroff
100 Bureau Drive
Gaithersburg
Maryland 20899-8940
ian.soboroff@nist.gov

Donna K. Harman
100 Bureau Drive
Gaithersburg
Maryland 20899-8940
donna.harman@nist.gov

National Institute of Standards and Technology

Abstract

While more and more people are relying on social media for news feeds, serious news consumers still resort to well-established news outlets for more accurate and in-depth reporting and analyses. They may also look for reports on related events that have happened before and other background information in order to better understand the event being reported. Many news outlets already create sidebars and embed hyperlinks to help news readers, often with manual efforts. Technologies in IR and NLP already exist to support those features, but standard test collections do not address the tasks of modern news consumption. To help advance such technologies and transfer them to news reporting, NIST, in partnership with the Washington Post, is starting a new TREC track in 2018 known as the News Track.

1 Motivation

News content has long been part of information retrieval test collections, but the search tasks that those collections measure is *ad hoc search*. Ad hoc search is a task where the user is seeking any and all information about a topic of interest. As such, articles are judged to be relevant to a topic if they mention the topic, even in a minimal way, as long as that mention is worth including in a report on the topic.

In 2018, people consume news overwhelmingly via social media recommendation, but also through web browsing, search, and advertising recommendation. Traditional news outlets more and more are taking a "digital first" strategy, rather than hewing to the notion of a newspaper front page. But the most change has come from social recommendation and news aggregators. Google News, started in 2002, marked the end of publisher-driven news delivery by pivoting the focus from the publisher to the story. The diversification of news delivery has democratized news publishing, and current news outlets reflect an enormous range of journalistic standards and methods.

NIST realized the time had come to reinvent news search as a focus for information retrieval and natural language processing research. In partnership with the Washington Post, NIST launched the News Track as part of the 2018 Text Retrieval Conference (TREC).[1] One component of this is a new document collection, the TREC Washington Post Collection, which is available as a free download from NIST. The second component is a pair of IR tasks driven by how content is structured for the Post's website.

2 Data

In partnership with the Washington Post, we have made a large archive of digital news content available to participants, extending from 2012 through August 2017. It contains both news articles and blogs as originally published by the Washington Post with a total of 608,180 documents (about 6.9GB uncompressed in size), divided into 12 text files. Each text file represents a collection of either news articles or blogs in one of those 6 years. The documents are stored in JSON format, with each line representing

In: D. Albakour, D. Corney, J. Gonzalo, M. Martinez, B. Poblete, A. Vlachos (eds.): Proceedings of the NewsIR'18 Workshop at ECIR, Grenoble, France, 26-March-2018, published at http://ceur-ws.org

[1] http://trec.nist.gov/

a single news or blog document. Each document has meta-data including article title, original article URL, author, date of publication, and sources for text and embedded media. For more information on how to obtain the TREC Washington Post collection, visit http://trec.nist.gov/data/wapost/.

We also have a reformatted dump of English Wikipedia from close to the time of the latest news articles available for download.

3 Tasks

On news outlets' websites, article content and hyperlinks are used to provide context and background. In other words, browsing is not arbitrary but is guided through stories in the sidebar and hyperlinks in the story to permit the reader to read more deeply. On the Washington Post's website, for example, related stories are manually linked both on the side and at the end of articles, and links within the article frequently link to related stories or further information about entities in the story.

However creating such links manually is a tedious and cost-ineffective process. It is not surprising that crucial background stories as previously reported or externally available are not always provided. Consider for example an article on February 4, 2018 titled "N. Korea to send nominal head of state to S. Korea". There is no single link to background information on the current state of the Korean conflict (other than one about Kim Jong Un's sister that was generated at the time of accessing this article under "Most Read World" dated *later* than the current article) , but there are no links to recent stories such as "Hot heads or cold feet? North Korea's mixed Olympic messages" and "North Korean athletes arrive in South Korea for Olympics" just reported a few days earlier, or "North Korea agrees to send athletes to Winter Olympics, South says" and "Vice President Pence will lead U.S. delegation to Olympics in South Korea" a month before. There was also a report back in 2014 about the North's high-level visit to the South at the end of the Asian Games, titled "North Korean officials pay rare and surprising visit to the South". Needless to say, many names mentioned in the current story have appeared in previous news articles and/or have entries in other online resources such as Wikipedia. If the journalist had had at his/her disposal a utility that can automatically retrieve those relevant stories in order of significance and link important entity mentions to more in-depth articles about them elsewhere, s/he would have been able to make them available to the reader with much ease.

Getting context to the reader is very difficult in the modern news landscape, but is more important than ever. IR and NLP technology can support journalists in suggesting links to articles and entities that provide background and promote a deeper understanding of a news story. For the first tasks in the News Track, we have chosen to work on background linking and entity ranking.

3.1 Tasks 1: Background Linking

The main task for this new track will be "Background Linking", defined as follows: given a news article, the system should retrieve other news articles that provide important context and/or background information that helps the reader better understand the query article. This task is essentially an ad hoc search with a specialized relevance criterion. Relevance for this task will be graded along a categorized scale:

0: the document provides **little or no** useful background or contextual information that would help the user understand the broader context of the query article.

1: the document provides **some** useful background ...

2: the document provides **significant** useful background ...

3: the document provides **essential** useful background ...

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

We will refine this category scale with the help of our partners at the Washington Post. The critical points are that relevance hinges on providing "useful background information or context", and that there are levels that align with utility for the reader. We anticipate that these relevance judgments would be made at NIST by NIST assessors, with training support from journalists and data scientists at the Washington Post.

As a research problem, we would like to investigate how this relevance criterion differs from "traditional topical relevance" both in how it is applied by assessors and how it measures systems differently. To that end, we may also ask whether the article is topically relevant to the query article. This could be implemented by adding one more level to the above scale to capture topical relevance.

We will use NDCG@5 [Jarvelin:2002] as the primary effectiveness measure: the sidebar has limited real estate, and should ideally contain the best contextualizing links. We will also report average precision and the other standard trec_eval measures.

The initial task is intentionally simple: we want to establish a baseline for the state of the art and use

that performance to consider refinements to the task. These might include:

- Having assessors cluster equivalent background articles, to allow the measure to support "retrieve one of these critical articles".
- Snipped generation for the sidebar, where the snippet should provide the critical context without the need to click through.
- Categories of background, for example about people and organizations. This would be measured using diversity metrics.

3.2 Task 2: Entity Ranking

The second task is "Entity Ranking": given an article, identify important entities mentioned in the article and rank those entities linkable to Wikipedia entries in the order of importance, in order to support the reader's understanding of the story. An example of an important entity might be "the Supreme Court", whereas an example of an unimportant entity might be "Washington" in a dateline.

By structuring this as a ranking task, we are separating out the core NLP problems of entity detection and linking from determining importance to the user. The provided mentions and links may be useful to researchers working on entity extraction as well.

Again working with criteria developed in conjunction with Post staff, we will identify the top entities in each article along a graded relevance scale, and measure the task as a retrieval task using nDCG.

4 SUMMARY

We set out the New Track with two initial tasks, Background Linking and Entity Ranking, which we believe are valuable to both the news creator and consumer. At the time of submitting this position paper, we are still in the process of refining the tasks and performance measurements via working with journalists and data scientists at the Washington Post and researchers in the IR and NLP communities. We welcome feedback and suggestions on the current tasks as well as recommendations on future tasks. We also encourage participation from researchers around the world.

4.0.1 Acknowledgements

Special thanks to Sam Han at the Washington Post for coordinating the efforts between the two organizations. We also appreciate the input from the other team members of the Retrieval Group at NIST.

References

[Jarvelin:2002] K Järvelin and J Kekäläinen. Cumulated Gain-based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, October 2002.