Dr. Dina Demner-Fushman
Behind the Mask

GB: Good morning. Today is May 28, 2021. My name is Gabrielle Barr, and I'm the archivist with the Office of NIH History and Stetten Museum, and today I have the pleasure of speaking with Dr. Dina Demner-Fushman. Dr. Demner-Fushman is a tenure track investigator with the Lister Hill National Center for Biomedical Communications at the National Library of Medicine (NLM), and today she is going to speak about some of her COVID work. So thank you very much for being with me.

My first question is what are some of the factors that led to the creation of the TREC-COVID, which stands for “text retrieval conference” evaluation, and when did you and others begin working on this project?

DDF: Good morning. When the pandemic started, the White House understood right away that the rapid development of the treatments and vaccines that we have now will not happen if the researchers do not know what is already available, do not have access to relevant scientific research through the literature, and the collection that is called CORD-19 [COVID-19 Open Research Dataset] was created. NLM had also a role in creation of that collection, and also of course the White House, and they reached out to industry. The Allen Institute for Artificial Intelligence was involved from the very beginning in creation of that large collection of research papers.

What distinguishes that collection from all the others is that the publishers were very open to providing access to everything that is published right away [about COVID-19] as opposed to other collections that usually include only peer-reviewed, more or less vetted, research. The collection started including preprints, everything that was published before peer review, right away. Then as soon as the collection was announced, all the information retrieval researchers—those who work on improving search engines such as you experience every day in your browsers when you look for something on Google—started looking into retrieving information from that collection. And what happened, they started reaching out to those of us who were already working in the biomedical domain, asking to evaluate the results of [the] systems.

Of course, it was not possible to evaluate the results of each one of those systems individually, so we got together, and we have a history of working together with the National Institute of Standards and with the Oregon Health & Science University and with The University of Texas Health Science Center at Houston (UTHealth). We form a team that usually organizes these evaluations on a regular schedule when there are no pandemics. We got together, and we said, “How can we organize an evaluation that will help the researchers as it all evolves, and what makes it different from what we usually know to do for these evaluations?” So that's how TREC-COVID came about.

GB: Yeah. What kind of information is included in the CORD-19 data set, and did you have any part in choosing what topics were included?

DDF: The CORD-19 collection of the scientific documents was [created] with the help of the medical librarians, and very broad searches were created to include everything. The orientation was to find everything that could somehow pertain to the topic. Some information about the previous coronavirus research could inform the SARS-CoV-2 research, so that's why the net was cast very broad, and then that's why we ended up with the very large collection to begin with. And then the fact that the COVID publications were prioritized led to the collection really shooting up in size with every round of our evaluation; with every release of the literature collection, it was growing. The topics came from various sources, and the topics were actually created by us, but these are still naturally occurring topics because they were informed by several sources. Some of the sources were the NIH researchers with whom we started talking, [asking] like what would be your information needs at this time, how can we help you? The same was happening at the Oregon Health and at UT Health. All of these medical libraries provided their searches, and we also looked at—because we, from the very beginning, thought that not only researchers and clinicians need information [but] policymakers need information, consumers need information. So we looked at the MedlinePlus logs to see what the consumers actually wanted to know, and that's how all these topics came together.

GB: Well, that's very interesting. How did you and others go about evaluating TREC-COVID?

DDF: You know the standard evaluation is relatively static. Usually, the evaluations are for determining which one of the approaches works better, how can we improve our search engines, and of course the goal of this evaluation was also how can we provide the best information possible in the context of a pandemic? And what makes the pandemic different, as we already discussed, the literature is uneven in quality and also the moment something goes from pre-print to print, you get several copies of that same publication. There's also a degree of change in that publication. So the AI2 [Allen Institute for Artificial Intelligence] team was trying to create these sort of folders: this is this paper in all its variants. So that's what makes it slightly different.

Then of course the growth. Usually the collection size is stable, but here the collection was growing all the time from round one, and we had five rounds of evaluations. From one round to another, the size of the collection was growing. Because the topics were evolving, we're also adding new topics, so the size of the search pool was growing as well, and the evaluation was done very rapidly, and again, we had volunteers from Oregon Health, UT Health, and the index section of the National Library of Medicine did a lot of evaluations last summer. The evaluation as usual is basically looking at two factors: Are you finding everything that you are supposed to find, and do you provide that relevant information up front?

GB: Were they pleased with the results, or was it mixed findings and you all have gone back and done some tweaking?

DDF: What we found is that the search engines are doing a pretty decent job on finding relevant documents, and we found that this collection was somewhat different because so much was published. It's unprecedented in size, of the human judgments on the quality of the retrieved documents. It's very large and still there are lots and lots of relevant documents, which is somewhat unusual, but it might stem from the fact that this collection to begin with was created to be relevant to COVID, so it was not a random universe or the whole universe. It was a subset that the broad queries were retrieving with already related searches, so that's why the task of re-ranking was so very important. So, yes, filter out the stuff that is not really relevant to that specific question but also from the large number of things that are relevant. Pick the ones that you need more upfront. That was the sort of the result that we know that the search engines can find all these documents, but we need the next step because the researchers particularly in that pandemic situation cannot sit there and read all these thousands of papers. That's why we had the sister evaluation that was called EPIC-QA for epidemic question answering, where instead of providing a relevant document, we actually try to provide an answer to the question.

GB: That's interesting. Can you talk more about EPIC-QA?

DDF: Yes. After TREC-COVID stopped and we had five rounds of evaluations with really, really good participation for TREC-COVID, and what was good [was] that teams from the big search engines participated. Also lots and lots of university teams. Some other institutions. I think overall we had close to 60 different organizations that participated in these five rounds.

GB: That's great.

DDF: Then when we stopped because we learned whatever there was to learn from this effort, we took the last document collection for the researchers, the CORD-19 that was used in the fifth round of the TREC-COVID evaluation, and then we took all the searches that we had developed for that collection, and we decided that in addition to answering professional or expert questions, we will take the same questions and try to answer them for consumers because consumers also want to know where did it come from, and consumers also wanted to know which kinds of masks are more useful than the others.

GB: Yeah.

DDF: So all of these things. But our hypothesis was that the answers will be different for the experts and the consumers. That's why we added consumer-friendly sources to the CORD-19 collection. We added CDC, World Health Organization, and whatever was published in some other sources. We added that information, and we created these topics, the description of what the information need might be. The questions were the same, but we created the description of information needs for consumers. We gave these to the other set of… well, we announced that here's another evaluation that's available. Then 12 teams participated in that first round, which was done to create training data for the question-answering systems. Then we created a set of new topics, and we had the main evaluation that could benefit from the learning, the training set that was created in the first round, and we had the second round of living evaluation. And in both of these evaluations we, of course, learned well, as we confirmed that when the current deep learning approaches have training data as in these five rounds of TREC-COVID and two rounds of EPIC-QA, we can see that when we have training data, the results get better.

GB: Interesting. What were some of the other things that you've learned from both of these projects?

DDF: That will kind of lead to another topic. So when we were talking to the index section about evaluating the answers, their main concern was that they will not be able to judge the quality of the answer: is it a good answer, is it a bad answer, or is it a misleading answer? We decided to split the task into not worrying about the quality of the answer and just sort of stating whether it is an answer or not. Then within the regular track cycle—so this text retrieval conference is close to 40 years in existence, and it's running on a yearly cycle, so research teams submit proposals for the research question that they want to study in that community-wide evaluation, and then there is a committee that reviews all these potential tracks within TREC and picks the ones that makes sense.

Within that regular cycle, there was a Misinformation track running for the second year and naturally they said, “We will focus on COVID misinformation.” So that made it very easy for us to say, “We're not going to worry about the quality of the answer; we are only worrying about if it is an answer or it is not.” And then within that TREC Misinformation track, there will be some different approaches to judge whether it's true or false or misleading. In that evaluation we actually just participated as a team, as one of the research teams, to contribute to these systems that try to distinguish what is true and what is not.

GB: That's very interesting. So what has been your particular role in all these initiatives?

DDF: Organizing, developing topics. Developing guidelines for judgments of the quality of retrieval judgments of the answers and then developing some methods for misinformation detection.

GB: It must be very time consuming to do all those things and very tedious. Can you talk about the process of how you go about it?

DDF: Tedious is when you are not interested in what you're doing. So time consuming, of course, yes, it takes a lot of time to do all these things. It takes a lot of time to do coding, but if you are into it, you actually do not notice how the time flies.

GB: That's good.

DDF: I have to say the Index Section, they contributed. Their time was limited by what was allowed because they were actually doing it within their regular work hours, and it was very nice of the NLM leadership to allow them to spend some of their time judging for these community evaluations.

GB: That's nice.

DDF: Some of the indexers—so for this EPIC-QA challenge—they had to develop answer keys. If you know the topic is “Where is COVID coming from?”, the answer key could contain items like “bats” and “pangolins” and “market” and “China.” They created these answer keys both for the expert level and for the consumer level, and some of them were so thorough and created like an 80 or 90 item list of what should be included in the answer.

GB: That's impressive. It's very impressive.

DDF: The majority of the answer keys are within like, maybe it's because it's the Index Section, it's like within 15 to 20 items on the answer key, but some are very, very long and exhaustive, and many aspects of the answer are required to be there.

GB: How do you go about selecting the new topics to include?

DDF: We were monitoring the logs and actually the discussions on social media. Whatever was the topic of interest that we did not have yet in our set, we would add that topic.

GB: That's interesting. Has there been other COVID research and initiatives that you've taken part in at NIH or outside of NIH?

DDF: As I said, we did the Misinformation TREC, which was focused on COVID. Some of the topics in the BioASQ challenge were also COVID-related, and we participated in that one as well. There are many, many COVID-related initiatives so you know someone will probably write a review on everything that was going on at the time. Yeah, it's only so many that you can participate in.

GB: Yes definitely. Well, have either the EPIC project or the TREC-COVID project—do you think that you would apply what you've learned with both of these resources to other kinds of things that you would do that are non-COVID related for the future?

DDF: Oh absolutely! You know whatever we, that TREC-COVID effort is the proof that whatever we do is very relevant to any emergency situation that might arise, and from whatever research we are doing, we learn something to advance. In our consumer health question answering system, we are implementing whatever we learned from the retrieval parts to find documents better, and also whatever we learned and implemented in our misinformation effort.

So we usually avoid dealing with lots of misinformation by including only trustworthy sources, right? But some years ago, we had a study that showed that for about 15% of the questions, these trustworthy sources do not have any answers, and you have to go out and search the web. Of course in that situation, it's better to say, "There is no reliable answer, and if you search the web, you might find these, but these are not reliable, these are not true," than to just give an answer to a person.

GB: That's interesting. Well, thank you very much for everything that you've been doing, and I wish you and those that you work with all the best success.

DDF: Thank you so much.