COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Jane Pan (Princeton University), September 22, 2021

Title: Contradiction Detection of COVID-19 Randomized Controlled Trials via BERT Language Models

YouTube Recording with Slides

September 2021 CIC Webinar Information

Transcript Editor: Macy Moujabber

Transcript

Lauren Close:

Slide 1

Quickly turn this over to our next speaker today Jane Pan. Jane is actually one of three winners of our inaugural CIC [COVID Information Commons] undergraduate student paper challenge. She's a first-place winner. The challenge was held earlier this spring, and we welcome Jane and are very excited to share her research with the broader CIC community. So, Jane, take it away.

Jane Pan:

Slide 2

Awesome, okay, let me share my screen real quick. Hopefully that's working. Not working? Okay, great awesome okay. So, I hope you've all been having a great start to fall. My name is Jane. I graduated from Columbia this past spring and I'm now at Princeton for graduate school. I'm very happy to present the work that I did during my senior year of undergrad with professor Chunhua Weng from Columbia University's Data Science Institute. Our project investigates contradiction detection of COVID-19 randomized controlled studies using mass language models such as BERT.

Slide 3

So, a little bit of context first. Contradictory results in clinical studies has been a long-standing problem for academics, researchers, and doctors alike, especially in a field such as high-volume publications. One study found that a third of original clinical studies are either challenged or unable to be replicated, and another found that a fourth or randomized controlled trials, in particular, are outright contradicted by

later findings. And this is an issue that became especially tangible during the outbreak of COVID-19. We've all heard about hydroxychloroquine whose initial clinical studies were really optimistic, and later the findings were contradicted pretty decisively. And so analyzing and interpreting results from a large and continually changing body of work is a challenge that's really important during time-sensitive scenarios like the global pandemic. So, to us, facilitating the process of identifying contradicting or agreeing studies would be really crucial for scientists who might want to, for instance, conduct systematic reviews, identify what might cause different results between two studies, evaluate the veracity of a research claim, and characterize the state of consensus or maturity on a particular research question. And so, for our research project, the question that we asked ourselves was how can we systematically extract evidence-based knowledge from raw text alone in order to quickly and automatically identify which studies agree and disagree?

Slide 4

So, we formulate this problem as a standard natural language inference, or NLI task, and the claim or the aim is to classify a pair of sentences as contradicting, entailing, or agreeing and neutral meaning that the sentence claims are unrelated or they neither entail nor contradict each other. So, the language model objective here is more formally given a pair of sentences x1 and x2 with some mass classification token CLS and some parameter matrix. We choose a label that maximizes the probability that the final state of CLS is that label for that specific x. And we choose mass language models here, specifically BERT, because these have historically had very strong performance of NLI tasks. We use pre-existing pre-trained models as the base model for our projects. The goal is to use transfer learning by adapting these models to our specific NLI task and we consider three base models. The first being the generic BERT model which is pre-trained on BookCorpus and Wikipedia and then two domain-specific models. BioBERT which is pre-trained on PubMed abstracts and articles and ClinicalBERT which is pre-trained on MIMIC 3 clinical notes.

Slide 5

So, for us the most crucial consideration was how fast the model would adapt to new research questions, because in practice you'd want the model to find contradictions in new research which it might not be pre-trained on. So, to that end we knew we needed a data set with research areas and questions that had never been seen before by the base models. And so, we made our own data set. We manually annotated a novel data set using LitCOVID, a publicly available database of COVID-19 PubMed articles. This is because COVID-19 is very recent and it wouldn't have been present in the data used to pre-train the base models. And in line with other biomedical analytic report annotation methods, we identified 15 separate research questions and 103 studies that answered them. So, we manually extract a sentence from each abstract that directly addresses the research question, and then two independent annotators manually labeled the pairs as contradiction, entailment, or neutral with respect to the research question. So, any labels that had disagreeing conclusions were tossed out.

Slide 6

To build our model, we add uninitialized classification layers to the base models and fine tune. We keep the base layer parameters frozen for now, since we have a relatively small train set and we only fine-tune the classification layers. For our train set we use ManConCorpus, a publicly available manually annotated medical inference NLI corpus very similar to what we did but more broad, not just COVID-19.

Slide 7

And we also reserve a small portion, about 20 of the LitCOVID data set for training. And we were really cautious about preventing contamination between test and train because the model has to generalize to questions that it hasn't seen before. So, to that end, we removed any pairs that mixed test and train sentences. So, what that means is that if a research question appeared in the train set, it will not appear in the test set and vice versa. We tokenize the sentence pairs, and for each base models we trained two models: one that added that small portion of LitCOVID data to his train and one that only used ManConCorpus. And we do this because we want to see how much the model improves with just a very small portion of COVID-19 specific data added to it.

Slide 8

So here are the results for our classification metrics of all the models, BioBERT and Clinical BERT with LitCOVID data performs the best, which makes sense because the base models are trained on a similar domain to LitCOVID, and unsurprisingly, if you add LitCOVID train data, it performs better than a model that doesn't. But I'd like to like note the improvement like the very drastic improvement with a very small proportion of COVID training data. So, the F scores are improving by a large degree. Almost all of them are like near doubling with the exception of the precision column, which is pretty strong across all the models.

Slide 9

Here we show the recall on a class by class basis and we see some pretty interesting patterns here. So far and away contradiction is a class that performs the best even with models that did not have any LitCOVID data training data added to it, and we hypothesize that this may be because negating terms like no or not are universal across topics and domains, so maybe the model can identify negations pretty quickly. We saw that LitCOVID training data primarily improves the neutral predictions. You can see it's like doubling the number of correct neutral predictions, which is most likely because it- now that it has LitCOVID training data it knows which COVID words don't necessarily, like, suggest contradiction or entailment. And entitlement is relatively weak overall except for BioBERT with LitCOVID, and we think that this might be because BioBERT pre-training corpora actually came from PubMed, so it may have learned features that helped to better identify textual affirmation or negation in a biomedical domain.

Slide 10

So, in summary, we have strong evidence to show that BERT models are a valid approach for conjugation detection in the biomedical domain. We have three pre-trained models which need only a small amount of training data for drastically improved performance. And just some error analysis very briefly. Some common patterns we found were, as you saw, earlier struggling with identifying mutual terms and then we saw some confusion with like abbreviations or medical terminology. So, for instance HCQ and hydroxychloroquine, the model doesn't immediately know they're the same thing so it'll say it's neutral or unrelated, and so on and so forth.

Slide 11

So, for the future, some of the interesting questions that we think could be answered are, like, how can we automatically select the best sentence without needing to manually extract this? And there are already some text nomination tools for clinical studies already in, like, openly available such as Trialstreamer or the Weng Labs picoparser, so I'd be interested in seeing how they could integrate this with a conjugation detection tool. And also, we're curious to know if we can improve the model's performance by, you know, supplying a user-provided list of acronyms or synonyms for that domain that the model is likely to come across.

That's all I have for today. I'd like to conclude by thanking Professor Weng and Dr. Hao Liu for their mentorship and help throughout my research and also a big thank you to Stan and Marguerite for their kind assistance with the project. Thank you for your time and I hope you all have a great week.