Week 3 - Supervisor Meeting Minutes

Meeting Date - 06/10/21
Meeting Location - SAWB
Meeting Attendees - Jake Lever and David O'Neill
Minutes Taken By - David O'Neill
Meeting Start Time - 14:30
Meeting End Time - 15:00

No agenda at this stage in the project so no particular structure to the meeting. Instead I have provided a narrative to the discussion.

The meeting began with an update from myself explaining the progress I have made in the last week. I discussed the fact that I had read over the TREC-COVID papers and worked on some pyterrier tutorials. The meeting continued with us discussing several key points identified in the research.

The CORD-19 dataset is very large and a high percentage of documents actually have no relevance to COVID. Jake discussed that the CoronaCentral dataset has filters on it to aim to have more relevant documents and also aggressively avoid duplicates. I mentioned that one of the successful TREC-COVID solutions makes use of date filtering, something which TREC-COVID does to keep the documents relevant.

Jake brought up the concept of altmetric, a statistic which is tracked for each document in the CoronaCentral dataset. It effectively gives the documents a score depending on how much they are mentioned publicly – for example on Twitter. He thought this would be a good way to potentially rank documents, as ones which are discussed more publicly are more likely to be relevant. He also thought it would be quite interesting to look at the altmetric scores for documents retrieved in TREC solutions, to see if this theory is proved to be correct.

We discussed the fact that it is probably best to retrieve documents by only referring to their title and abstract, since there are legal implications involved with accessing the rest of the papers. The current search tool just uses keyword matching on the titles, since some documents don't have abstracts. Jake acknowledged that it might be a good idea to look at matching keywords related to the topic tagged for the document in the dataset, as even a retrieval technique using something like BM25 would probably produce similar results to keyword matching on the title alone.

Another point raised by Jake was the idea of using different types of indexing in order to do a better job in the first round of retrieval. It seems that this is the fundamental area which we can try to improve upon the current CoronaCentral search tool.

Jake also pointed out that it might be a good idea to look at the user logs for the website and consider improving them so that we can evaluate how well the current search tool is working. We could also implement some form of A/B testing to compare it to the new tool we create.

The plan for the next couple of weeks is to continue researching and noting down key areas to discuss. Afterwards, we can come up with a concrete plan and deliverables for the project.

Week 4 - Supervisor Meeting Minutes

Meeting Date - 13/10/21

Meeting Location - SAWB

Meeting Attendees - Jake Lever and David O'Neill

Minutes Taken By - David O'Neill

Meeting Start Time - 14:30

Meeting End Time - 15:00

No agenda at this stage in the project so no particular structure to the meeting. Instead I have provided a narrative to the discussion.

The meeting began with an update from myself explaining the progress I have made in the last week. I explained that I had completed the Terrier tutorials and that had given me some experience adopting doc2query and DeepCT. I talked about the altmetric scores and how I had read some literature to familiarise myself with them. I discussed some possible second stage strategies such as Bert or sequence to sequence models. Finally I explained that I had been generally experimenting with the CoronaCentral dataset to find out what I can do with it.

Jake then proposed two ideas which he had thought of throughout the week. He explained how most of the documents have generic boilerplates at the beginning of the them, explaining the origins of Covid-19 or something to that effect. Obviously, this background information is not relevant to the actual topic of the document, so is it possible that taking it out would be beneficial? Would we be able to devise a strategy to identify generic, irrelevant information and remove it? Also, it might be something that DeepCT picks up on – does it think that this information is important or does it give it a small score for relevance? It might be a good idea to look at TREC-COVID false positives for this.

Secondly, he thought it would be interesting to look at date filtering. If we filter our corpus by date relative to when the disease started (such as COVID-19 after 2019) then how many documents will be removed and are they actually relevant? Might be something to experiment with.

I pointed out that I had found a pretrained Bert model on CORD-19 – we discussed that this could be very useful or possibly not, especially since there is a lot in CORD-19 which is irrelevant. We therefore spoke about the alternatives to Bert such as SciBert and PubMed Bert, to understand their differences and see how the might suit the project.

Finally, we discussed mapping the CoronaCentral documents back to their associated CORD-19 documents, in order to find out the queries which were judged to be relevant to them and also to identify their altmetric scores. This can be done by looking at the CORDUID field in both the datasets and matching them. There could be problems as CORD-19 has many duplicates which have different IDs.

We agreed that over the next week the plan was to keep researching before coming up with a more solid plan. Areas to experiment on include removing irrelevant background information from documents, analysing date filtering and looking at mapping CoronaCentral documents back to CORD-19 documents.

Week 5 - Supervisor Meeting Minutes

Meeting Date - 20/10/21

Meeting Location - SAWB

Meeting Attendees - Jake Lever and David O'Neill

Minutes Taken By - David O'Neill

Meeting Start Time - 14:30

Meeting End Time - 15:00

No agenda at this stage in the project so no particular structure to the meeting. Instead I have provided a narrative to the discussion.

We went through all of the areas I have been researching in the past week. First we discussed the date filtering graph which I had sent before the meeting. Startlingly, it shows that the majority of documents published before 2000 were classed as relevant by TREC-COVID. Jake agreed that this requires some deeper analysis to dive into the dataset and see what is happening. It could be the annotations made by the TREC judges, it could be the dataset. Only further exploration will tell. Jake suggested looking into the documents prior to 2000, and looking at the queries they were classed relevant for. It may be that there is a query which is relevant prior to 2000. If it seems there is a flaw in the dataset, I decided to have a look into some literature, to find out whether there has been any professional critiquing of CORD-19.

Next we discussed how I had been taking all the papers that I have read and trying to form a narrative about them in order to write a literature review.

I also updated Jake on the fact that I had been mapping the CoronaCentral dataset back to the CORD-19 data for evaluation purposes.

We then moved onto a discussion about how to remove generic, "boilerplate" background information on covid. The clustering technique seems like a good idea in practice but definitely will require refinement to pick the correct training data and cluster. It seems the current clustering is not specific enough. Jake suggested taking the word covid out of every abstract and seeing what results came from that.

Finally, Jake thought it would be good to split the dataset up into training, validation and test for now to avoid a mess later on.

We agreed that the format of a status report and minutes is working well and that no formal agenda is required for now, as basing meetings as informal chats seems to be working.

Week 6 - Supervisor Meeting Minutes

Meeting Date - 27/10/21

Meeting Location - SAWB

Meeting Attendees - Jake Lever and David O'Neill

Minutes Taken By - David O'Neill

Meeting Start Time - 14:30

Meeting End Time - 15:00

No agenda at this stage in the project so no particular structure to the meeting. Instead I have provided a narrative to the discussion.

As usual, the meeting was structured by going through each item on the status report.

We discussed the date anomaly throughout both CORD-19 and CoronaCentral. I believe that it occurs because the queries which are looking for coronavirus are bringing up old documents, whereas the queries which are looking for COVID-19 are bringing up new documents. This might not be particularly relevant, because nowadays, if someone is looking for information on coronavirus, they probably mean COVID-19, not SARS or MERS for example. It might be a good idea to look at some data visualisations for queries, showing their average, max and min publish years, for example.

We discussed splitting the data into sections for training, validating and testing. I had split the documents up into sections, but should also consider how I am going to split up the queries.

We talked about the clustering experiment and how it seemed like a good idea in practice but did not improve results in reality. It could be because we became too focussed on one query, and need to think more about the bigger picture as a whole.

I also mentioned that I had found some good papers on critiquing CORD-19 which mention how altmetrics could improve the dataset. This is promising for the future of the project.

Jake finished by mentioning how he believes the project could go. I will use pyterrier to make perform some experiments on the standard ways of querying CORD-19 and CoronaCentral, taking inspiration from the strategies used in TREC-COVID. If there is one that stands out as a winner, it can be put on CoronaCentral. We can also go off on tangents along the way by thinking up new unusual ways to tweak the retrieval method and experiment to see how they work.

Week 7 - Supervisor Meeting Minutes

Meeting Date - 03/11/21

Meeting Location - Virtual

Meeting Attendees - Jake Lever and David O'Neill

Minutes Taken By - David O'Neill

Meeting Start Time - 14:30

Meeting End Time - 14:50

No agenda at this stage in the project so no particular structure to the meeting. Instead I have provided a narrative to the discussion.

As usual, the meeting was structured by going through each item on the status report.

We began by discussing the publication dates for documents classed as relevant for particular queries. I had made an error the following week, but now the problem is cleared up, it is evident that the more recent the document, the more likely it is to be classed as relevant. This provides me with several options for building the retrieval system and also backs up the same consensus in literature.

We then examined some violin plots for each query, digging into when the associated documents were published. Jake described this as a "sanity check" as it did not really provide much information, but instead backed up our current thoughts. It appears as well that there is the possibility that queries for the first round have older documents associated with them. This is not so much the case for queries at subsequent rounds as there was a newer collection of papers to choose from. This led to a discussion about the judging criteria and the actual judges themselves, with the takeaway being that it is always important to consider that there might be flaws in the annotations.

I then discussed how I split up the queries into train, validation and test sets and began doing some experimentation using typical TREC solutions. TF_IDF was the best document representation. I tried some features and trained them, with coordinate ascent having the best outcome. Then I tried some neural reranking which was unsuccessful due to the fact that most of the training crashed my colab sessions. Then the neural reranking took 6 hours before crashing again. We therefore discussed

certain ideas about how to approach this - maybe filtering out documents before indexing which did not have a high altmetric value or had an old publication date? Jake also suggested looking into some literature behind non neural indexing, so identifying what was there before doc2query. Maybe this would be a quicker and more efficient form of indexing.

Essentially, the takeaway from the meeting was that I should keep on experimenting, trying all sorts of methods to identify strategies which work and rule our ones which don't.

Week 8 - Supervisor Meeting Minutes

Meeting Date - 10/11/21

Meeting Location - Virtual

Meeting Attendees - Jake Lever and David O'Neill

Minutes Taken By - David O'Neill

Meeting Start Time - 14:30

Meeting End Time - 15:00

No agenda at this stage in the project so no particular structure to the meeting. Instead I have provided a narrative to the discussion.

We began by examining the diagrams I provided prior to the meeting. We went through every feature in the CoronaCentral dataset, to look at the values which were attached to relevant documents. Jake suggested that it would be a good idea to compare these to diagrams which illustrate what the data says about irrelevant documents. For example, comparing the number of relevant documents with a DOI to the total distribution papers with a DOI overall.

We then discussed a brief plan for the next week. We will both delve into some research to find out what other indexing methods are available, apart from Doc2Query and DeepCT. It seems that lots of machine learning models are being applied because it seems like the best method, even though it might not be applicable to our dataset and problem domain.

I will also create a notebook which performs some retrieval, using the most successful method I have found so far. We will sample around 20 false positives and 20 false negatives and examine them in close detail. This is because it might show us if any particular problems are cropping up or whether they are all caused by different problems. Additionally, it will allow us to evaluate the annotation process. We always need to remember that the documents were annotated by humans, and so there could be an element of error.

The meetings will be conducted in person from next week onwards following the end of COP26.

Week 9 - Supervisor Meeting Minutes

Meeting Date - 17/11/21

Meeting Location - SAWB

Meeting Attendees - Jake Lever and David O'Neill

Minutes Taken By - David O'Neill

Meeting Start Time - 14:30

Meeting End Time - 15:00

No agenda at this stage in the project so no particular structure to the meeting. Instead I have provided a narrative to the discussion.

We first discussed the best strategy I have uncovered so far - filtering the dataset by the presence of an altmetric score, indexing the title and abstract then creating TF_IDF embeddings. So far it performs the best and I have been using it as the benchmark for all my experiments.

We first looked at the notebook which looks at some mistakes which were made in the retrieval. We discussed that there are some annotations which seem a bit incorrect and so Jake is going to read into them further. I also should look into the annotation process to understand why some mistakes might have been made. I also mentioned that the papers classed as molecular biology tend to cause errors so Jake is going to look into the classification of these papers.

We looked at some comparison graphs to show the characteristics of relevant/irrelevant documents and found that there really wasn't too much difference. Perhaps the topics and journals could be trained as features for the first stage of retrieval.

We then looked at the NDCG scores for each TREC query and saw that the poor performing queries were the vague, generic ones like "Coronavirus outside the body". From this, I mentioned that it might be a good idea to look at the current coronacentral searches so Jake is going to have a look at the logs to find out what people are searching for. He is also going to look into whether the users are using the autocomplete tool, for example searching for a drug and clicking to view the page for that drug when the automated search is generated.

We also looked at the current coronacentral search tool and saw that it doesn't perform very well at all. From this, it is clear that my new tool is definitely an improvement. Jake said that my next step would be to build a search api for this. I should turn it into a restful api which could be deployed somewhere like AWS, maybe using something like Lucene.

Week 10 - Supervisor Meeting Minutes

Meeting Date - 24/11/21

Meeting Location - SAWB

Meeting Attendees - Jake Lever and David O'Neill

Minutes Taken By - David O'Neill

Meeting Start Time - 14:30

Meeting End Time - 15:00

No agenda at this stage in the project so no particular structure to the meeting. Instead I have provided a narrative to the discussion.

As usual, we addressed each point in my status report. First, we discussed the coronacentral logs. My visualisations highlight the most viewed pages on coronacentral which led to some observations. First, the queries are all very technical, they are not vague like some of the TREC-covid queries. Secondly, Jake noted that some searches were binary, with the user looking for more than one topic. So an example might be "covid origin and covid spread". I will look into how my search tool performs when given binary queries. Jake also noted that he does not have complete confidence in the accuracy of the logs, as it might be sometimes logging 1 view as multiple views. As such, it is important to use this data with caution.

Next we looked at comparing the coronacentral queries with the TREC queries. From this, Jake expressed that it would be a good idea to roughly classify each of the queries, to effectively be able to compare them. For example, it might be a good idea to say "30% of the queries were about drugs in coronacentral logs, but only 20% of TREC queries were about drugs". We are effectively looking for a way to be able to compare the TREC queries with real-life queries. If I can say that my tool isn't getting perfect results on TREC but is getting good results on real-world queries, then we will have something good to say.

We then looked at the documents returned by popular coronacentral queries and observed that they are sensible results and give us confidence that our tool is performing well.

Finally, I demonstrated the search API that I intend to use for the queries. At the moment it is hosted locally by running a colab cell with flask ngrok, but in the future I will work on configuring my virtual environment to run it from the command line. I will also keep looking into using Lucene although that might cause more bother than it's worth.

Week 11 - Supervisor Meeting Minutes

Meeting Date - 1/12/21

Meeting Location - SAWB

Meeting Attendees - Jake Lever and David O'Neill

Minutes Taken By - David O'Neill

Meeting Start Time - 14:30

Meeting End Time - 15:00

No agenda at this stage in the project so no particular structure to the meeting. Instead I have provided a narrative to the discussion.

We began by discussing how I classified the real-life coronacentral queries. Jake agreed that while this was not a perfect process, it was a good estimate of the query classifications given the information I have. We analysed the classifications, noticing that therapeutics is a very sought after topic, while certain topics which are included in the TREC queries are not seen in the real world at all. Jake mentioned that this is an interesting discussion point and with some deeper thinking, it could form the basis of a short paper.

I then demonstrated to Jake my search API and the associated front-end. I explained that it is formed by building an index on colab which is filtered by the presence of an altmetric score, with the most likely relevant documents expanded with deepct and then embeddings created with TF_IDF and reranked by altmetric score. I explained that this is not the easiest thing to evaluate, so Jake mentioned the possibility of being able to set it up on the coronacentral site at the beginning of next year and putting some logging in place which could see how real-life users are experiencing it.

Finally, we talked about the binary searches in coronacentral. It is clear that some of these searches all appear to be of the same topic and are likely from the same person. I said that I will continue working on the binary searches, looking into how possibly some neural indexing might affect it? Or if it might be better to take "and" out of queries or perform two queries and intersect the results?

This was the last meeting before the Christmas holidays and we agreed that the best process was for me to continue working by myself and communicating by email if necessary.

Week 17 - Supervisor Meeting Minutes

Meeting Date - 12/1/22

Meeting Location - Online

Meeting Attendees - Jake Lever and David O'Neill

Minutes Taken By - David O'Neill

Meeting Start Time - 14:30

Meeting End Time - 15:00

No agenda at this stage in the project so no particular structure to the meeting. Instead I have provided a narrative to the discussion.

We began by discussing my plan for going forward, looking into how the dissertation will be structured and what the content will be. Jake is happy with the plan and noted that this will hopefully give us an idea of any gaps in research.

Next we discussed a deadline for finishing all practical work. I have suggested the end of January in order to give myself plenty of time to write up the dissertation. Obviously if any new obstacles arrive, I can go back to practical work after this deadline.

Then we talked about the possibility of ethics approval. Jake stated that the analysis of logs provided to me so far does not require approval as the data was not gathered for the purpose of my project. If, however, we go onto analyse logs concerning the new IR tool, this would require approval.

In terms of access logs, Jake is going to have a look at the google analytics of CoronaCentral in order to establish what metrics can be gathered regarding user sessions. This should give us a way of evaluating the current and the new tool. In the case that this is not possible, I suggested one other way to evaluate could be a pop up which asks the user whether the search was useful to them.

Finally we discussed the technical details relating to the integration of the new tool into the website. My plan is to experiment with different ways of indexing the dataset and running the API. This may be something which takes longer without the use of a Colab GPU, so it will be important to establish an appropriate strategy. Jake will look into the

hosting side of this, to understand how we can best deploy the API on the AWS instance.

Week 18 - Supervisor Meeting Minutes

Meeting Date - 19/1/22

Meeting Location - Online

Meeting Attendees - Jake Lever and David O'Neill

Minutes Taken By - David O'Neill

Meeting Start Time - 14:30

Meeting End Time - 15:00

No agenda at this stage in the project so no particular structure to the meeting. Instead I have provided a narrative to the discussion.

We discussed that I have been working on a plan for the dissertation, which has identified some gaps in research, specifically the evaluation of the new IR tool versus the one currently in CoronaCentral. We discussed the way that I could evaluate them - typically an NDCG score would be best and how I can go about determining this for the CoronaCentral tool. Then we had a look at precision and recall, understanding how the new tool improves upon the old one. Jake emphasised that these are good metrics to evaluate, but to make clear in the dissertation what they mean and what they tell us about the solution.

Then we discussed whether there is the potential of leveraging the topic types that CoronaCentral has in order to improve the tool. Jake recommended using the deep learning tool that he has developed, available on hugging face, in order to classify the queries made by the user. Depending on the topic which the query is classified as, this might be used to produce more accurate results. If for example, we have classified the query to be about long covid, then we can try and put an emphasis on the documents which are tagged as long covid.

Finally we discussed the configuration of the API, agreeing that the two file format is a good way to go for integrating the tool into CoronaCentral.

My plan is to perform accurate evaluations of the tools with ndcg scoring, working on leveraging the deep learning data and finally continue planning and beginning to write the dissertation.

Week 19 - Supervisor Meeting Minutes

Meeting Date - 26/1/22

Meeting Location - Online

Meeting Attendees - Jake Lever and David O'Neill

Minutes Taken By - David O'Neill

Meeting Start Time - 14:30

Meeting End Time - 15:00

After a slight change in plan regarding the content of the dissertation, most of the discussion centred around some experiments that I carried out and our plans going forward. I discussed the baseline experiments that I am working on, followed by the four types of experiment that I would like to carry out using CoronaBERT.

We then discussed the plans for the dissertation, deciding that the design and implementation chapters may be slightly overlapping and so could be condensed into one.

The plan for the following week is to have my baseline experiments fully implemented and some sound evaluations carried out. I will then aim to work on the new experiments, notably beginning by performing hyper parameter tuning on the model I have created. By next week, I hope to have some concrete results to show Jake.

Week 20 - Supervisor Meeting Minutes

Meeting Date - 02/02/22

Meeting Location - Online

Meeting Attendees - Jake Lever and David O'Neill

Minutes Taken By - David O'Neill

Meeting Start Time - 14:30

Meeting End Time - 15:00

I began by talking through my status report for the week. I have spent time tidying up all my previous work and concretely showing it in scientific experiments. I talked Jake through the results that I have obtained - I will do a sanity check to make sure whether BM25 can really be improved upon that much by reranking.

Next I discussed the process of beginning the dissertation and that I am now on the design chapter. I have decided it would be good to have separate design and implementation chapters, with the former focussing on high level ideas and the latter focussing on the actual experimental process.

Next we discussed the correct referencing process for tools I am using - I will make reference to them both in the notebook in which they are being used and the dissertation.

Finally, we discussed the appropriateness of our test data. We are unsure whether we are recreating the final round of TREC-COVID, so we will continue to research and understand whether this is the correct strategy.

Week 23 - Supervisor Meeting Minutes

Meeting Date - 25/02/22

Meeting Location - Online

Meeting Attendees - Jake Lever and David O'Neill

Minutes Taken By - David O'Neill

Meeting Start Time - 13:30

Meeting End Time - 14:00

I began by discussing the work which I have carried out over the past couple of weeks. Firstly, I talked through the work I did on by making use of Altmetrics, filtering retrieved documents by presence of Altmetric score and also reranking by an Altmetric regression model. I also explained that I had spent some time working on the presentation of my notebooks, ensuring that they are tidy, readable and runnable. As such, I have ensured that all necessary datasets are available on my public github repository for downloading.

I then went on to discuss the structure of my dissertation, explaining to Jake the specific chapters and sections which will be included and what they will outline. Jake noted that he is able to offer simple guidance with the dissertation, simply highlighting whether the format and grammar/tone are acceptable.

We agreed to meet at the same time the following week, due to the ongoing UCU strike action.

Week 24 - Supervisor Meeting Minutes

Meeting Date - 4/03/22

Meeting Location - Online

Meeting Attendees - Jake Lever and David O'Neill

Minutes Taken By - David O'Neill

Meeting Start Time - 13:30

Meeting End Time - 14:00

This was a very quick meeting. I simply updated Jake on the progress I had made for the previous week which included preparing my repo for submission and also working through my dissertation further.

I have a couple of areas which I think need improvement. Notably, the implementation and evaluation section seems to be quite convoluted in areas. Additionally, the final discussion section on the CoronaCentral dataset could potentially be covered in too much depth and so I aim to fix this.

Jake is kindly going to read over my first draft briefly to give some very high level structural and stylistic pointers.

Week 25 - Supervisor Meeting Minutes

Meeting Date - 9/03/22

Meeting Location - Online

Meeting Attendees - Jake Lever and David O'Neill

Minutes Taken By - David O'Neill

Meeting Start Time - 14:30

Meeting End Time - 14:40

I had very little to report this week as there was only a short window of time between this meeting and the last one. I explained that I have been working on editing my dissertation, making my presentation and tidying up my repository for submission.

Jake noted one key stylistic point when writing the dissertation - to ensure that every paragraph begins with a topic sentence. I hope that this will assist me in tidying up the final chapters, as I still feel they are somewhat convoluted in places. This should help me get my point across concisely.

Week 26 - Supervisor Meeting Minutes

Meeting Date - 16/03/22

Meeting Location - Online

Meeting Attendees - Jake Lever and David O'Neill

Minutes Taken By - David O'Neill

Meeting Start Time - 14:30

Meeting End Time - 14:45

This was a very short meeting with only a very small amount of discussion required. We discussed some stylistic points about the dissertation to clear up a few small queries that I had regarding formatting. We agreed that I have achieved a sufficient amount of content and my task is now to continue refining the dissertation.

Due to the industrial action, there will be no meeting next week. As I am unavailable the week beginning 28th March due to work commitments, Jake assured me that it is absolutely fine to submit a week early. I have always set myself the deadline of 25th March and the fact that the deadline was extended has not changed that for me. As such, I intend to submit on 25th March and this will be our last meeting.