November 22 - December 6 Summary

In these two weeks, my work involved training the BioFLAIR classifier. At this point, I've given up on using BioBERT. I trained the model on the `bc5cdr` dataset for 81 epochs (it stopped at 81 epochs because the Early Stopping criteria were met). Training took 6h 12m on the provided machine (NVIDIA T4 GPU, 16GB RAM). I incorporated the BioFLAIR model, adapted to our needs, in the pipeline code. The model works well in finding medical terms, and can also handle multi-word terms.

Below is a sample run of the pipeline, where I run some pre-processing steps as described in previous reports, picking one sentence that seemed to have several medical terms.

```
Python 3.6.9 (default, Nov  7 2019, 10:44:02)
[GCC 8.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> from drugdiscovery.preprocessing import *
>>> from drugdiscovery.postprocessing import *
>>> from drugdiscovery.classification import *
// A bunch of warnings
>>>
>>> pruner = LowConfidencePruner('../../../newJSON')
>>> pruned = pruner.prune()
>>>
>>> term_remover = IrrelevantTermRemover(pruned)
>>> filtered = term_remover.remove()
>>>
>>> test = [filtered[4]]
>>> test
["rise blood pressure and blood sugar one of the investigators cardiologist outer Nygard commented the very high intake of total and saturated fat did not increase the calculated risk of cardiovascular diseases or Omeprazole are the most popular treatments for heartburn and reflux disease the FDA consider such so safe that it has approved like Nexium Prevacid and Prilosec for over-the-counter sale but new side effects are still being discovered by researchers and lies medical records of more than $244,000 don't they found that those taking ppis were 19% more likely to experience a stroke over the following six years the higher the dose the greater the risk this was an epidemiological study so cause and effect if not been established the investigator suggest a follow-up clinical"]
>>>
>>>
>>> classifier = FlairClassifier(save_dir='../../../BioFLAIR/ner/models/v1/')
>>> classifier.classify(test)
2019-12-03 20:40:13,083 loading file ../../../BioFLAIR/ner/models/v1/best-model.pt
[['cardiovascular diseases', 'Omeprazole', 'heartburn', 'reflux disease', 'Nexium Prevacid', 'Prilosec', 'stroke']]
```

There is a little bit of work left--specifically, the duplicate term remover doesn't work as well as I had hoped, so that needs to be fixed. This shouldn't take longer than a day to fix.

Thoughts

- It seems that for this particular sample, the Google Cloud transcriber didn't do such a great job. For example, I think there's a word missing after "cardiovascular diseases", and that the sentence ends there. I think it should be, "[drug] or Omeprazole are the most popular ...".
- Forming pairs of adjacent terms ignores the "[drug] treats [disease 1] or [disease 2] construct". I never thought of this before. Still, the NER that we did is a necessary step.
    - I think this could be fixed if we had a dataset that not only classified medical terms as "Entities" as `bc5cdr` does, but as drugs and diseases. There are two other datasets provided by the BioFLAIR authors, but these only have diseases.
    - So I wonder if we could train two models--the first is the one above that finds both, and the second being one that only finds diseases, and then performing a set difference to get the drugs. We would need to maintain the order somehow and then form adjacent pairs accordingly. This wouldn't take very long either--maybe around 12 hours on that machine.

Future work

- Fix de-duper
- Use some metrics to check how well we're doing
- Run the two-classifier model to see if that improves the performance
- Update the final report document