Week 3 Milestones and Progress Report

Timothy Laurent

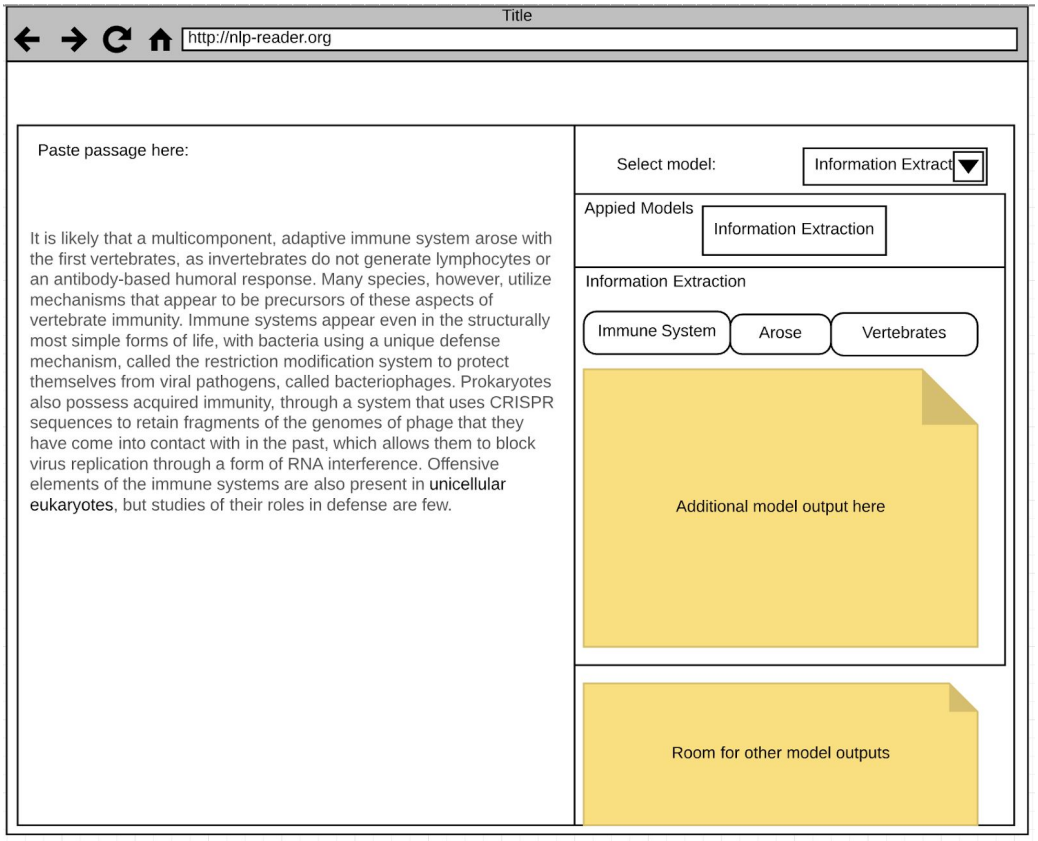
Project description:
A tool to present different NLP-derived projections of textual data to facilitate understanding. The hope is that this tool can consume different models that the user can select. Initially there were planned several modules such as word simplification, clause simplification, open information extraction, and summarization. Eventually it would be nice for the user to be able to give feedback to the models though the UI but that will be saved for a later effort.

NLP Investigation
Models that were initially thought to be easy-to-use published methods, proved to be more domain-specific, for example with Biomedical Text Simplification or computationally intractable and slow as is seen with for Wikipedia summarization tool. The former uses a synonym dictionary of MESH terms for text simplification making it ungeneralizable and the latter requires intensive training on multiple GPU setups which wouldn't work for a real-time app.

Areas where I had success were with information extraction, I was able to pull out triples of information using pyClauseIE and using the OpenIE model from AllenNLP. The main issue with this was there was no coreference resolution, so it's pulling out pronouns. As a next step for a viable tool I'm going to combine the triples with coreference resolution to generate a more useful information representation where all references to the same entities collapse into a single representation. Showing this to the users will be the first NLP-based visualization and this model output will be used in a survey to guage the model's utility.

Although I didn't have luck with using the Tensor2Tensor Transformer model, I'm going to look into using pretrained Transformer-based BERT model to do summarization instead. These models have had great success in many tasks in NLP so I will see if we can harness them into a realtime app. There is a new BERT model that is trained on PubMed and PubMed Central Open Access in addition to Wikipedia that I'll evaluate for this task.

UI Wireframe



The UI will be a simple design a left panel where text can be pasted and a right panel for model selection and viewing. On the right there a select widget where the model can be selected; below that, an area displaying all selected models; and below that panels of model output.

Development

I have a basic react app set up and some code to serve AllenNLP Models using a Flask backend. The front end will send raw text and an array of models to the backend which will pass these texts to models and pass the annotated text back to the front end. The frontend will have a react component to parse and display the content of each model type.

The output from the models contains character offsets to the original text so highlighting on mouse hovers should be possible.

A service based architecture will be used where each model exists as a service that the application can call, this will decouple the models from the flask app that is handing the requests from the front-end.

Survey reading tasks:

These questions were selected from the SQUAD 2 dataset in the immune_system set. SQUAD 2 has questions that do not have an answer making a more difficult task. These examples contain complex sentences and domain specific terms and concepts. This task, I think, will be challenging enough for humans to result in a lower baseline thus an effective intervention to facilitate comprehension will be more noticeable in the data. I'm still investigating options for performing the survey considering both the in-class option and using Mechanical Turk.

In addition to the selected reading comprehension question 2 question to assess the confidence and enjoyment will be asked.

Question 1.

It is likely that a multicomponent, adaptive immune system arose with the first vertebrates, as invertebrates do not generate lymphocytes or an antibody-based humoral response. Many species, however, utilize mechanisms that appear to be precursors of these aspects of vertebrate immunity. Immune systems appear even in the structurally most simple forms of life, with bacteria using a unique defense mechanism, called the restriction modification system to protect themselves from viral pathogens, called bacteriophages. Prokaryotes also possess acquired immunity, through a system that uses CRISPR sequences to retain fragments of the genomes of phage that they have come into contact with in the past, which allows them to block virus replication through a form of RNA interference. Offensive elements of the immune systems are also present in unicellular eukaryotes, but studies of their roles in defense are few.

Invertebrates do not generate what type of cells that are a part of the vertebrate adaptive immune system?

- *Ground Truth Answers:* lymphocytes lymphocytes lymphocytes

What is the main defense mechanism of bacteria known as?

- *Ground Truth Answers:* the restriction modification system restriction modification system restriction modification system

The restriction modification system is used by bacteria for protection from what pathogens?

- *Ground Truth Answers:* bacteriophages viral bacteriophages

What is the system by which prokaryotes retain phage gene fragments that they have previously come in contact with?

- *Ground Truth Answers:* CRISPR CRISPR sequences CRISPR

Where is it unlikely that the first multicomponent, adaptive immune system arose?

- *Ground Truth Answers:* `<No Answer>`

What arose in the latest vertebrates?

- *Ground Truth Answers:* <No Answer>

What mechanisms do many species not utilize?

- *Ground Truth Answers:* <No Answer>

What kind of immunity do prokaryotes not have?

- *Ground Truth Answers:* <No Answer>

What has often been studied in unicellular eukaryotes?

- *Ground Truth Answers:* <No Answer>

Question 2.

When a T-cell encounters a foreign pathogen, it extends a vitamin D receptor. This is essentially a signaling device that allows the T-cell to bind to the active form of vitamin D, the steroid hormone calcitriol. T-cells have a symbiotic relationship with vitamin D. Not only does the T-cell extend a vitamin D receptor, in essence asking to bind to the steroid hormone version of vitamin D, calcitriol, but the T-cell expresses the gene CYP27B1, which is the gene responsible for converting the pre-hormone version of vitamin D, calcidiol into the steroid hormone version, calcitriol. Only after binding to calcitriol can T-cells perform their intended function. Other immune system cells that are known to express CYP27B1 and thus activate vitamin D calcidiol, are dendritic cells, keratinocytes and macrophages.

What does a T cell extend when it encounters a foreign pathogen?

- *Ground Truth Answers:* a vitamin D recepto rextends a vitamin D receptor vitamin D receptor

What is the active form of vitamin D known as?

- *Ground Truth Answers:* calcitriol calcitriol steroid hormone calcitriol

What is the nature of the relationship between T-cells and vitamin D?

- *Ground Truth Answers:* symbiotic relationship symbiotic symbiotic relationship

What gene is responsible for converting calcidiol into calcitriol?

- *Ground Truth Answers:* gene CYP27B1 CYP27B1 gene CYP27B1

Other than T cells, what other immune cells express CYP27B1?

- *Ground Truth Answers:* dendritic cells, keratinocytes and macrophages dendritic cells, keratinocytes and macrophages dendritic cells

What does a pathogen do when it encounters a T-cell?

- *Ground Truth Answers:* <No Answer>

What cell binds to vitamin C?

- *Ground Truth Answers:* `<No Answer>`

What vitamin do T-cells have a parasitic relationship with?

- *Ground Truth Answers:* `<No Answer>`

What gene is expressed by the B-cell?

- *Ground Truth Answers:* `<No Answer>`

What gene converts calcitriol into calcidiol?

- *Ground Truth Answers:* `<No Answer>`

Additional Questions:
On an 1-5 scale from strongly disagree to strongly agree, rate your agreement with these statements.

I feel confident about my answers.

I enjoyed this quiz.