Problem Definition

• Given a medical trial report and a structured question prompt about the findings in the report, find the evidence and answer the question.
• There is a need to automate this, given the volume of biomedical evidence being released everyday.
• The size of these reports (up to thousands of words) causes significant computational challenges.
• Dataset – Annotated Pub-med articles and prompts (10k).
• Prior approaches don’t use transfer learning, evidence attention is scattered.

Ideal Approach

Neural Network

vector

attention

vector

Some Text Encoder

article

intervention

comparator

outcome

Our Approach

• We use a pipelined approach to handle processing the entire article through a BERT encoder.
• We first train an evidence classifier to pick the possible evidence sentences.
• We then train a model to predict the answer based on the top evidence sentences.

Results and Findings

• The evidence classifier is trained using pairwise positive and negative samples. The negative samples include random non-evidence sentences, as well as evidence sentences for other prompts.
• We experimented with hinge loss with margins in the range (0.5, 0.8), and binary cross entropy loss.
• We later switched to using paragraphs instead of sentences to include more context.

Evidence Classifier

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SciBERT + Hinge</td>
<td>84.5</td>
</tr>
<tr>
<td>SciBERT + BCE</td>
<td>81.2</td>
</tr>
</tbody>
</table>

Predictor (oracle)

<table>
<thead>
<tr>
<th>Model</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours - SciBERT</td>
<td>81.6</td>
</tr>
<tr>
<td>Previous best</td>
<td>73.9</td>
</tr>
</tbody>
</table>

Heuristics

<table>
<thead>
<tr>
<th>Model</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heuristics</td>
<td>43.1</td>
</tr>
<tr>
<td>Previous best</td>
<td>53.1</td>
</tr>
<tr>
<td>Ours – SciBERT Pipeline</td>
<td>42.2</td>
</tr>
</tbody>
</table>

Conclusions & Next Steps

• The evidence classifier and predictor seem to be working well on their own. But the overall pipeline is broken. This needs to be fixed.
• A lot of sentences lack context. The marked evidence sentences don’t have the prompt words in them. We tried to add some context but they are still lacking.
• We will try to bridge the two models in the pipeline using a weak signal - something on the lines of negatively reinforcing evidence sentences that don’t help the downstream predictor.