Evidence Selection in Long Documents

Subendhu Rongali, Rajarshi Das
University of Massachusetts Amherst

UMassAmherst
College of Information & Computer Sciences

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.


Running title • August 2012 • Vol. XXI, No. 1

Article Title

JOHN SMITH*
University of California
john@smith.com

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

I. INTRODUCTION

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer nec odio. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

II. METHODS

Maecenas sed ultricies felis. Sed imperdiet dictum arcu a egestas.

- Donec dolor arcu, rutrum id molestie in, viverra sed diam
- Curabitur feugiat
- turpis sed auctor facilisis
- arcu eros accumsan lorem, at posuere mi diam sit amet tortor
- Fusce fermentum, mi sit amet euismod rutrum
- sem lorem molestie diam, iaculis aliquet sapien tortor non nisi
- Pellentesque bibendum pretium aliquet

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultri-

*A thank you or further information

1

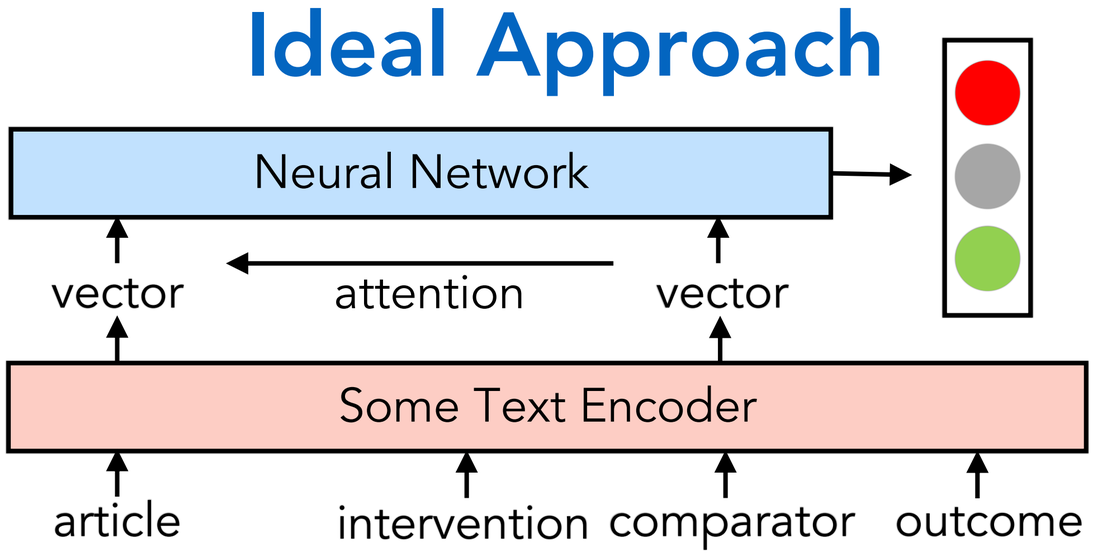
Intervention: *metronidazole*
Comparator: *placebo*
Outcome: *pre-term birth*

prompt

Patients receiving metronidazole experienced significantly fewer pre-term births than those in the comparison group.

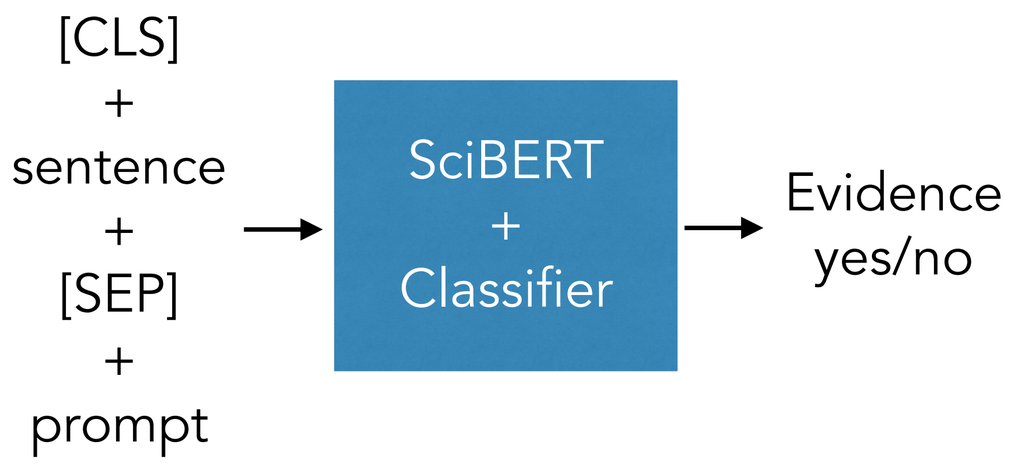

sig. decrease | no sig. difference | sig. increase

Ideal Approach

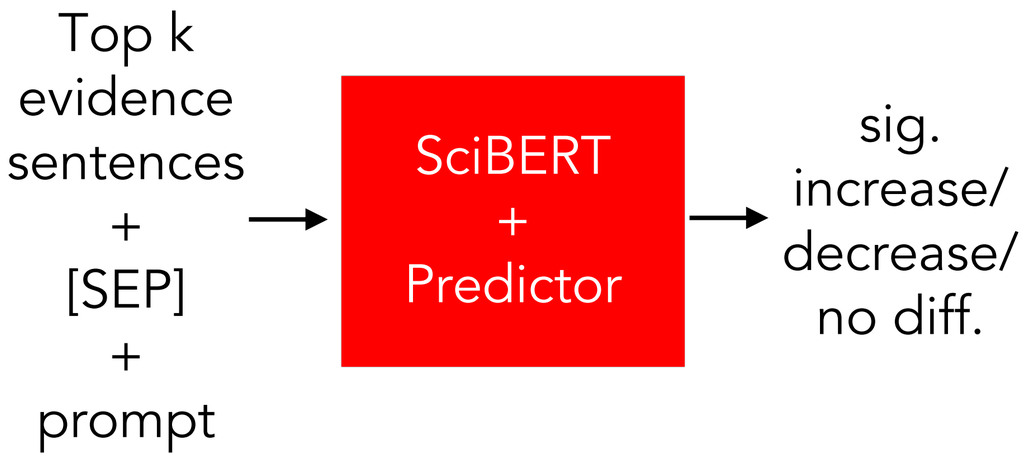


Our Approach

- 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.



- We train by sampling positive and negative evidence sentences and maximizing their margin.
- We then run this model on all sentence to get their evidence scores and pick the top k.



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	
	Accuracy
SciBERT + Hinge	**84.5**
SciBERT + BCE	81.2
Predictor (oracle)	
	F-score
Ours - SciBERT	**81.6**
Previous best	73.9
Overall	
Heuristics	43.1
Previous best	**53.1**
Ours – SciBERT Pipeline	42.2

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.