SuMe: A Dataset Towards Summarizing Biomedical Mechanisms

Mohaddeseh Bastan, Nishant Shankar, Mihai Surdeanu, and Niranjan Balasubramanian
Contact: mbastan@cs.stonybrook.edu

PubMed has indexed more than 1 million publications per year in the past 8 years!

we need applications that can organize biomedical information.

Task Definition

Create a dataset and task which can read biomedical abstracts and find the relation between entities and explain the reason behind the underlying relation.

Example:
What is the relation between *cadmium* and *potassium?*
Explain the reason.

Given:

- Sentences from a scientific abstract
- Pair of entities in the abstract *(cadmium, potassium)*

The kidney is a main target organ for <el> *cadmium* <le> toxicity. The present study has been performed to test for effects of <el> *cadmium* <le> on electrical properties of cultured subconfluent kidney (MDCK) cells. <el> *Cadmium* <le> leads to a rapid, sustained and reversible hyperpolarization of the cell membrane, paralleled by an increase of the <re> *potassium* <er> selectivity and a decrease of the resistance. Thus, <el> *cadmium* <le> increases the <re> *potassium* <er> conductance of the cell membrane. The half maximal effect is elicited congruent to 0.2 microM, a concentration encountered during chronic <el> *cadmium* <le> intoxication. At extracellular calcium concentration reduced to less than 0.1 microM, 5 microM <el> *cadmium* <le> leads to a transient hyperpolarization, which can be elicited only once. High concentrations (50 microM) of <el> *cadmium* <le> lead to a sustained hyperpolarization even at extracellular calcium concentrations of less than 0.1 microM. According to fluorescence measurements <el> *cadmium* <le> leads to an increase of intracellular calcium activity, which is sustained at 1 mM and transient at less than 1 microM extracellular calcium activity. <end>

Generate:

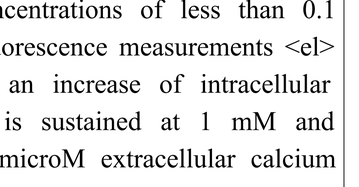
- Relation that connects the entities:
  - Positive activation *OR*
  - Negative activation
- Sentence explaining the mechanism behind the relation:
  - Why is the relation true? *OR*
  - How does the relation come about?

<exp> In conclusion, <el> *cadmium* <le> at low concentrations enhances the <re> *potassium* <er> conductance in a calcium dependent way. positive. <end>

Previous Work

Biomedical NLP:

- Extracting Information
- Organizing Text
- Summarizing Text



Explanation NLP:

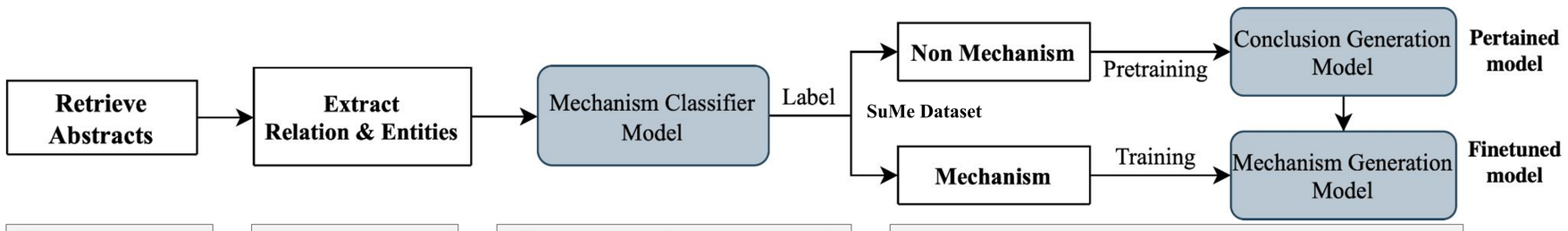
- Relation Extraction with explanation
- Question Answering with explanation

Our work focuses on explaining the mechanism underlying a relation from supporting sentences in scientific literature.

How to create this dataset

- Understanding scientific literature needs experts
- Hard to create a large scale dataset:
  - Time consuming and Expensive
- Instead, we ask experts to identifying mechanisms, train a small classifier, use that to create a large scale dataset

Data Collection



-Extract abstracts from PubMed
-Filter-out abstracts with no conclusion sentence

Run REACH to identify the main entities and relation between entities

- Collect a set of human annotated data (mechanism/non mechanism)
- Train a binary classifier
- Label all conclusion sentences with the trained classifier

-Use non-mechanism sentences for pretraining the conclusion generation model
-Use mechanism sentence for training the mechanism generation model

Data Statistics

Dataset	Train	Dev	Test
Abstracts	20765	1000	1000
Avg. #words in conc.	33.7	34.9	33.5
Avg. #sent. in supp.	12.15	12.44	12.33
#Unique entities	12685	1357	1364

Data Quality

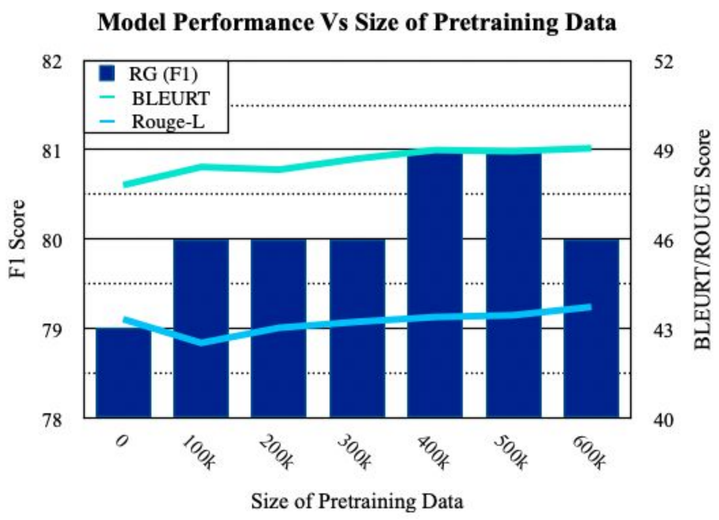
Five biomedical experts evaluated 125 mechanism sentences:

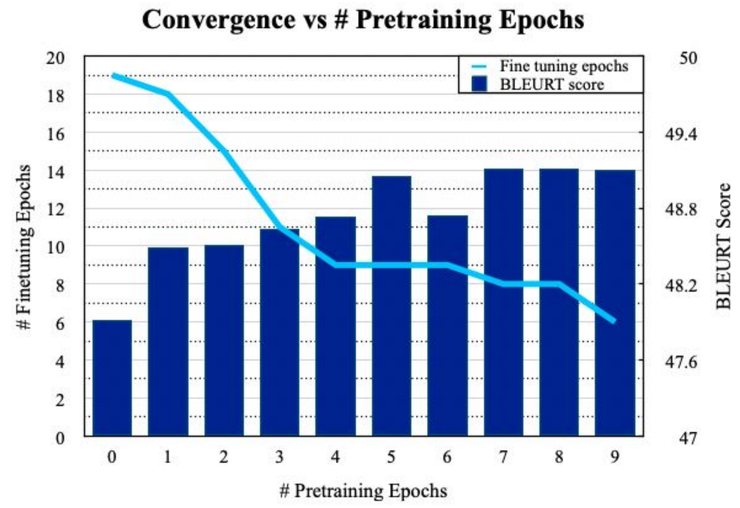
Quality	Correct
Entities & Relation Extraction	90%
Mechanism Sentence Classifier	85%
Instances w/o noise	84%

Baseline Models

Model	Relation Generation (F1)	BLEURT
BART	76	42.49
GPT2	74	44.19
T5	72	44.41
GPT2-PubMed	78	46.33
SciFive	**79**	**47.81**

Effect of Pretraining





Error Analysis

Manually categorized the errors in 100 outputs with worst BLEURT:

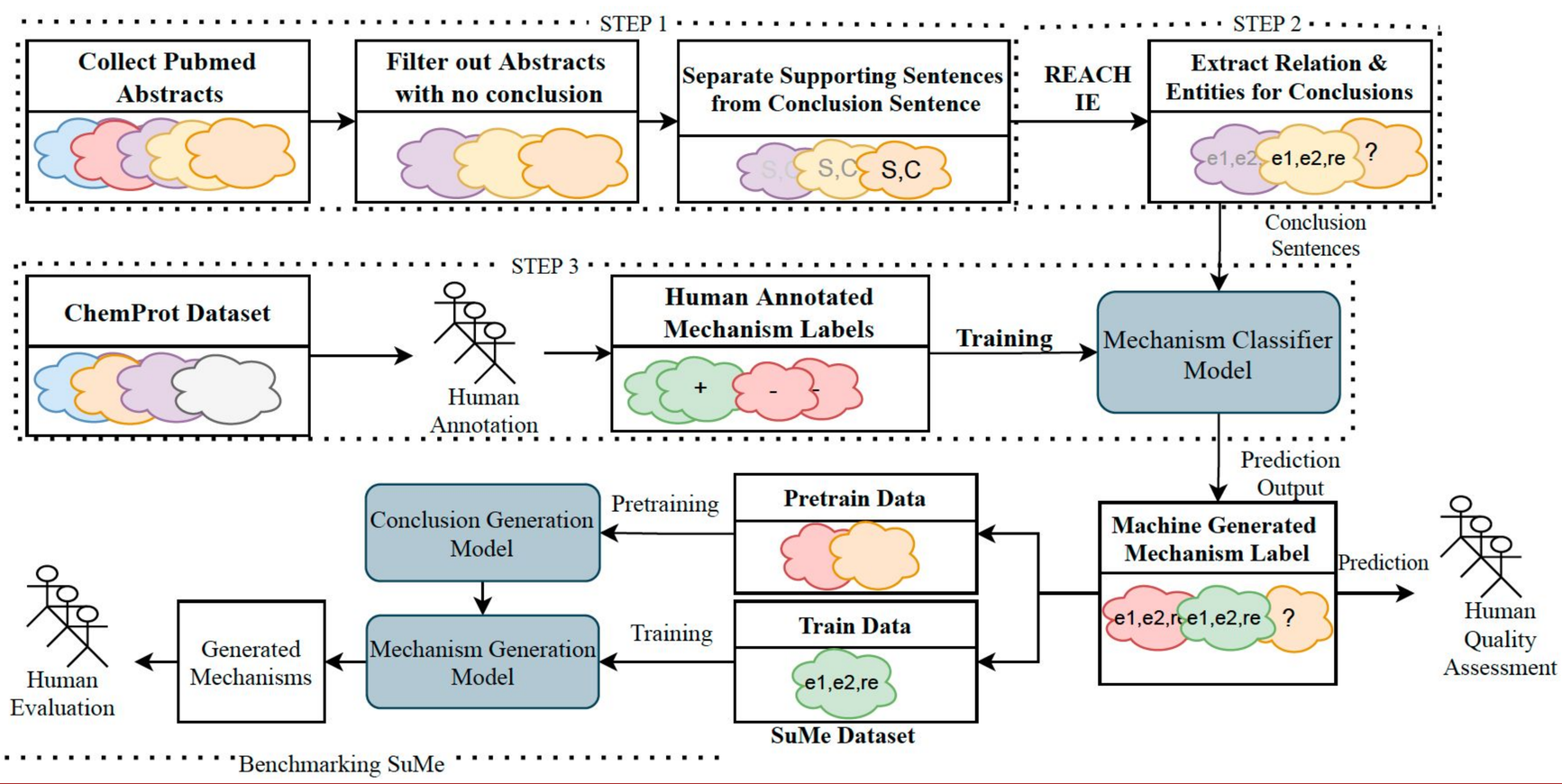
- Missing Entities (35%)
- Incorrect Mechanism (24%)
- Flipped Relation (19%)
- Non Mechanisms (11%)
- Multiple pieces of information (11%)

Generation Example

Ground Truth	Generated Example
In conclusion, the increase in *SNGFR* caused by *ANF* is associated with an increase in glomerular capillary hydraulic pressure and with a blunted maximal tubuloglomerular feedback response.	We conclude that *ANF* inhibits maximal tubuloglomerular feedback by increasing glomerular capillary hydraulic pressure and by decreasing *SNGFR*.
We conclude that, without modulatory factors which play a role in vivo, *NGF* can enhance the synthesis of *tyrosine* hydroxylase n sympathetic ganglia in vitro, provided organ culture conditions which permit optimal survival of adrenergic neurons are selected.	This conclusion was further supported by pulse-labeling of *tyrosine* hydroxylase with [3H]leucine, which showed that *NGF* increased synthesis of *tyrosine* in sympathetic ganglia by approximately 50%.

Blue: Expected mechanism, Green: good generated mechanism, Red: wrong generated mechanism

Summary of the Paper



In Conclusion, this paper introduces

- New Task and Dataset: Mechanism Summarization and SuMe
- A semi-automatic bootstrapping process yields large scale training (and pre-training) data for the task
- Benchmarking and analysis show many challenges remain for biomedical text understanding

SuMe Dataset and Code

Please visit our website:

https://stonybrooknlp.github.io/SuMe/