Clinical Literature Entity-Relation Extraction using Pre-trained Language Models

HNSC 7200 Seminar
Oct 19 2021

Yoonsik Park (University of Manitoba), Dr. Serena Jeblee (University of Toronto),
Dr. Noah Crampton (University of Toronto, Mutuo Health)

Background

Objectives

Dataset

Methods

Results

Background

Objectives

Dataset

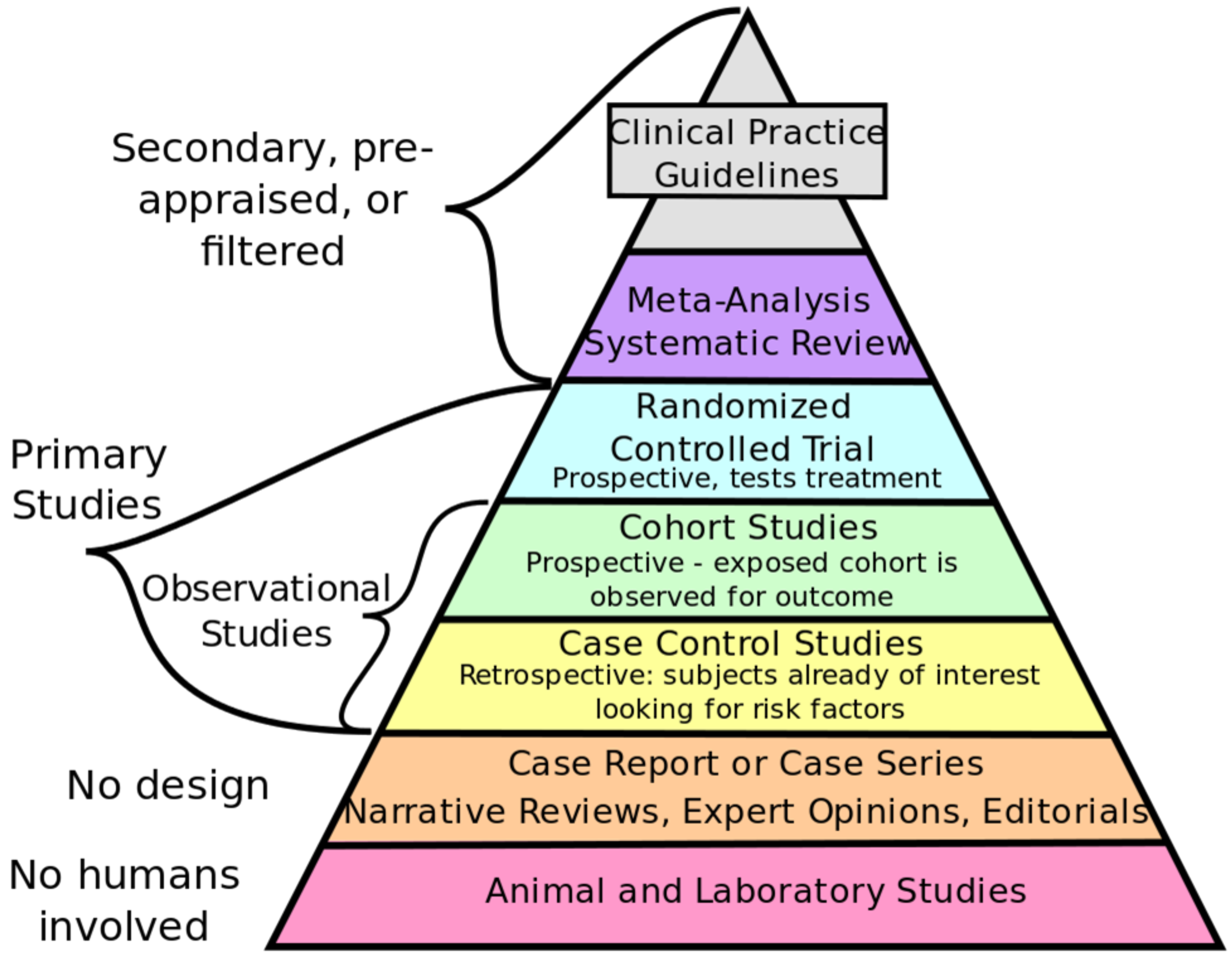
Methods

Results

Literature Review and Meta-Analysis is Essential

Background

- Systematic Literature Review is considered **Gold Standard**
  - PICO framework is commonly used
- If results from primary studies are statistically combined, it is considered a **Meta-analysis**
  - Increased power, decreased bias, stronger conclusions
- Basis for evidence-based medicine

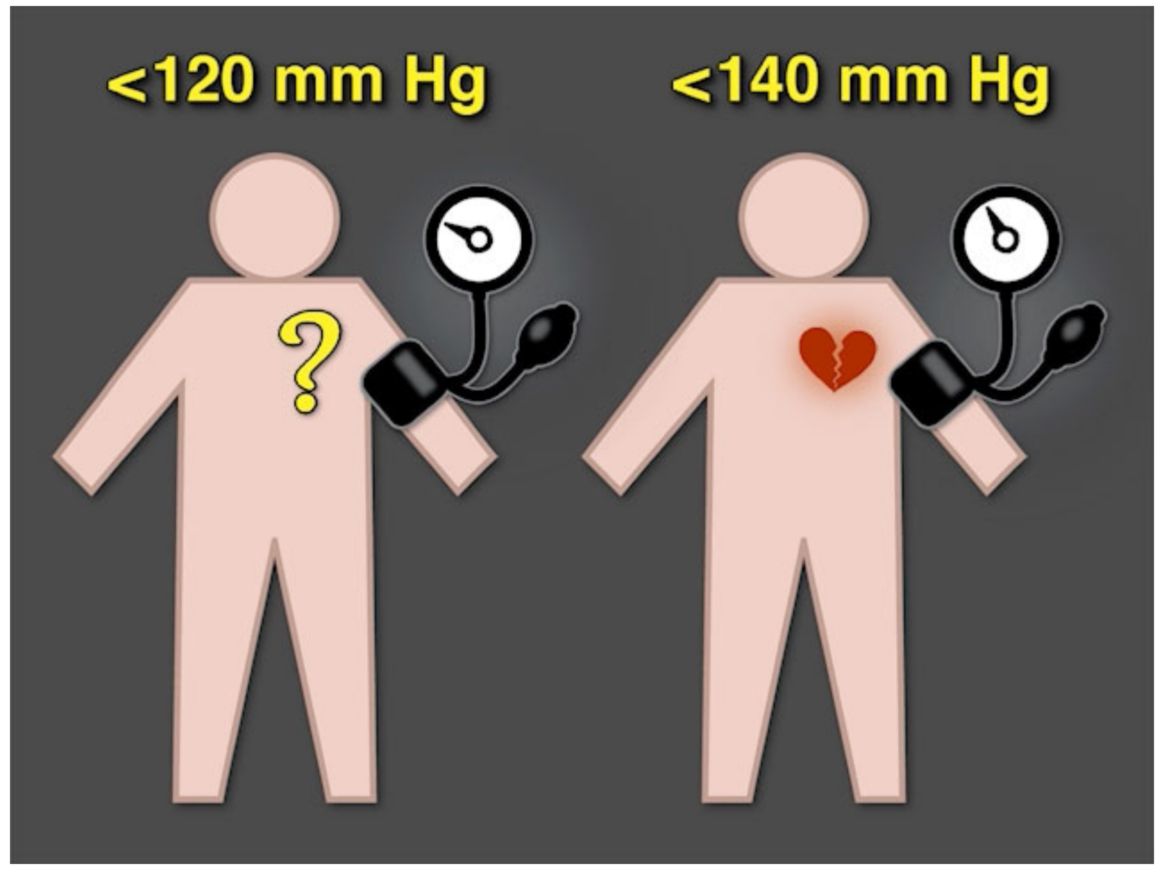


PICO Framework Helps us Understand Clinical Trials

Background

Example Trial: The SPRINT, Intensive vs. standard blood pressure control (2016)

- **P**opulation: Patients aged 75 years or older
- **I**ntervention: Systolic Blood Pressure <u><120 mmHg</u>
- **C**omparator: Systolic Blood Pressure <u><140 mmHg</u>
- **O**utcome (Primary): Major Adverse Cardiac Events (MACE)
  - **Result**: Hazard Ratio (HR) **0.66**
- **O**utcome (Secondary): All-cause Mortality
  - **Result**: Hazard Ratio (HR) **0.67**



Systematic Reviews are Well-defined and Structured

Background

- Steps include: search strategy, selection criteria, quantitative/qualitative analysis, summary
- Some steps require a human to read article text and extract key terms, a few examples:
  - **Applying selection criteria:** date of publication, geographical location, cohort characteristics (age, sex, etc.), methodology
  - **Study quality appraisal:** methodology, sources of bias
  - **Data collection:** risk estimates (ratios), outcome, explanatory, control variables


European Journal of Epidemiology
https://doi.org/10.1007/s10654-019-00576-5

GUIDELINES

Check for updates

A 24-step guide on how to design, conduct, and successfully publish a systematic review and meta-analysis in medical research

Taulant Muka[1] · Marija Glisic[1,2] · Jelena Milic[3,4] · Sanne Verhoog[1] · Julia Bohlius[1] · Wichor Bramer[5] · Rajiv Chowdhury[6] · Oscar H. Franco[1]

Received: 21 June 2019 / Accepted: 29 October 2019

Abstract
To inform evidence-based practice in health care, guidelines and policies require accurate identification, collation, and

Machine Learning is Rapidly Improving

Background

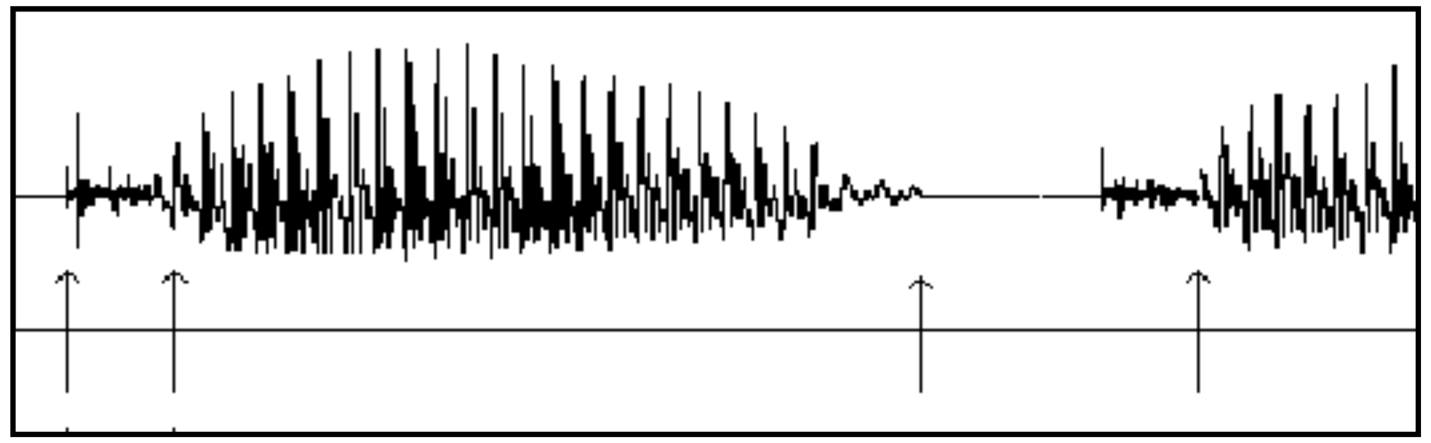
- ML can be loosely defined as "training" computer algorithms to automatically perform a task that we desire
- A growing subfield of ML is Natural Language Processing (NLP), i.e. the field of processing and analyzing large amounts of natural language data (text)

Image Classification

Cat

Dog

Speech Recognition



PyTorch

TensorFlow

NLP is a Versatile Tool

Background

- Sentence Classification

The room was great, but the staff were unfriendly → Negative Sentiment

- Named Entity Recognition (NER) **

Stephen Hawking (Person) was a physicist (Occupation) who was born in Oxford (Location)

- Relation Extraction (RE) **

The patient received 20mg (Dosage) of metoprolol (Drug) , b.i.d. (Freq.)

Biomedical NLP is Showing Success

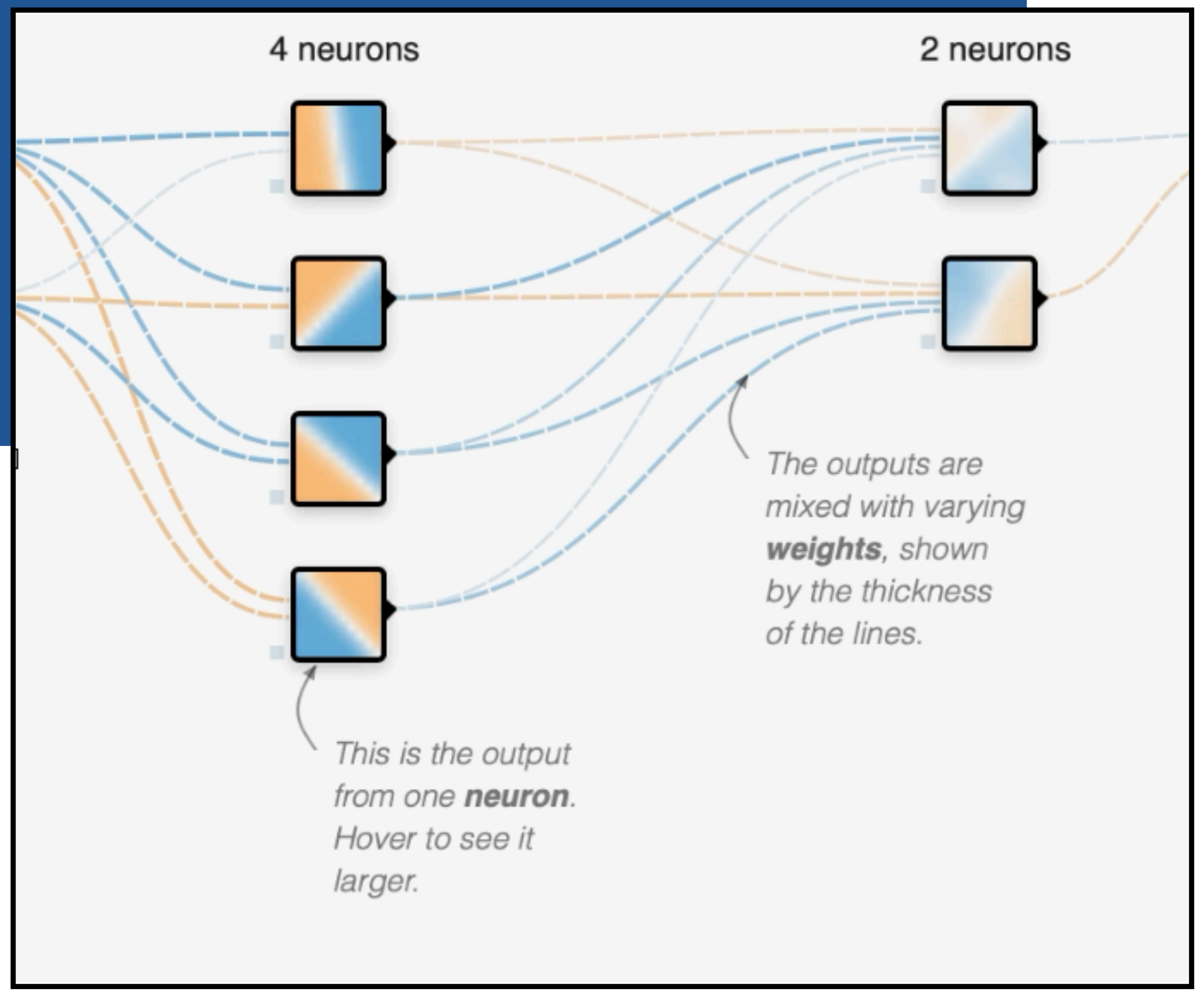
Background

Biomedical NLP is the process of extracting and synthesizing important information from biomedical data

- Notably, Schmidt *et al.* (2020) showed that pre-trained language models were able to support systematic reviews by
  - classifying each sentence as: Population, Intervention, Comparator, or Outcome (PICO)
  - predicting a single span in a sentence that best answers a PICO question
- Limitation: a sentence may contain information from all PICO categories & multiple sections from a sentence may be needed to answer a PICO question

Rationale

Background



- Clinical systematic reviews are the gold-standard for evidence based medicine
- However, they are time consuming, requiring researchers, planning, and money
- 1.4 million articles released on PUBMED in 2019! We need a way to scale this process
- The structured format of a systematic review implies that we can automate parts of this process using biomedical NLP

Primary Research Objective

Objectives

Extract risk estimates (Hazard, Risk, Odds Ratio) with their linked study variables (Explanatory, Baseline, Outcome) using an NLP model that can extract multiple entity-relations from a sentence.

In terms of PICO: Explanatory → Intervention, Baseline → Comparator

Secondary Research Objective

Objectives

Explore different variations of model to improve performance:

- Combine NER model and RE model into a single Joint Entity-Relation Extraction model
- Try different language model variations (size + corpus)
- Add context, i.e. the title + first sentence from an abstract

Background

Objectives

Dataset

Methods

Results

PubMed Abstracts

Dataset

- 79,468 PubMed abstracts published from Nov 11 2019 to Dec 31 2019 were downloaded and filtered for risk estimates → pool of 6,421 articles
  - Training set: 82 articles → 95 sentences with risk estimates
  - Test set: 31 random articles → 32 sentences with risk estimates
- Annotated entities: `"RR", "OR", "HR", "Outcome", "Baseline", "Explanatory"`
- Relations between the risk estimates and study variables were tagged

Training - Example

Dataset

Rates of MACE (RR 0.91 95% CI 0.58-1.41; p = 0.66; I2 75%), MI (RR 1.75 95% CI 0.87-3.55; p = 0.12; I2 0%) and ischemic stroke (RR 0.83 95% CI 0.53-1.31; p = 0.42; I2 0%) did not differ between the OAC monotherapy and the OAC combination therapy.

PMID: 31748954, Title: Antithrombotic therapy for chronic coronary syndrome and atrial fibrillation: less might be more.

Training - NER Task

Dataset

Outcome RR Outcome RR

Rates of MACE (RR 0.91 95% CI 0.58-1.41; p = 0.66; I2 75%), MI (RR 1.75 95% CI

Outcome RR

0.87-3.55; p = 0.12; I2 0%) and ischemic stroke (RR 0.83 95% CI 0.53-1.31; p = 0.42;

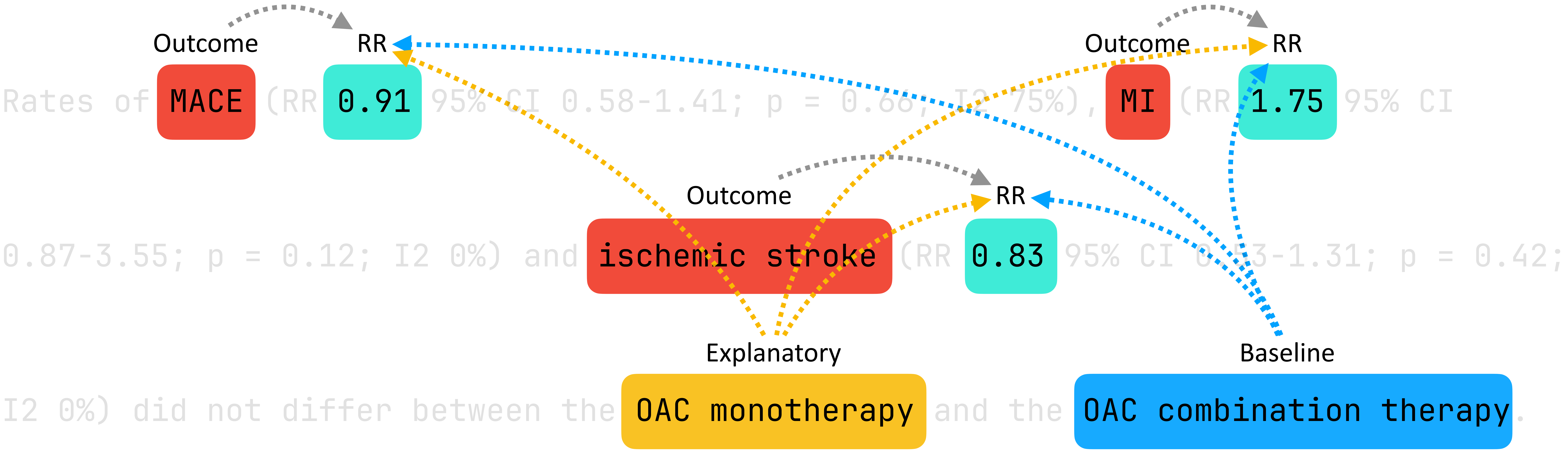
Explanatory Baseline

I2 0%) did not differ between the OAC monotherapy and the OAC combination therapy.

PMID: 31748954, Title: Antithrombotic therapy for chronic coronary syndrome and atrial fibrillation: less might be more.

Training - Relation Task

Dataset



PMID: 31748954, Title: Antithrombotic therapy for chronic coronary syndrome and atrial fibrillation: less might be more.

Background

Objectives

Dataset

Methods

Results

Pre-trained Language Models are a Starting Point

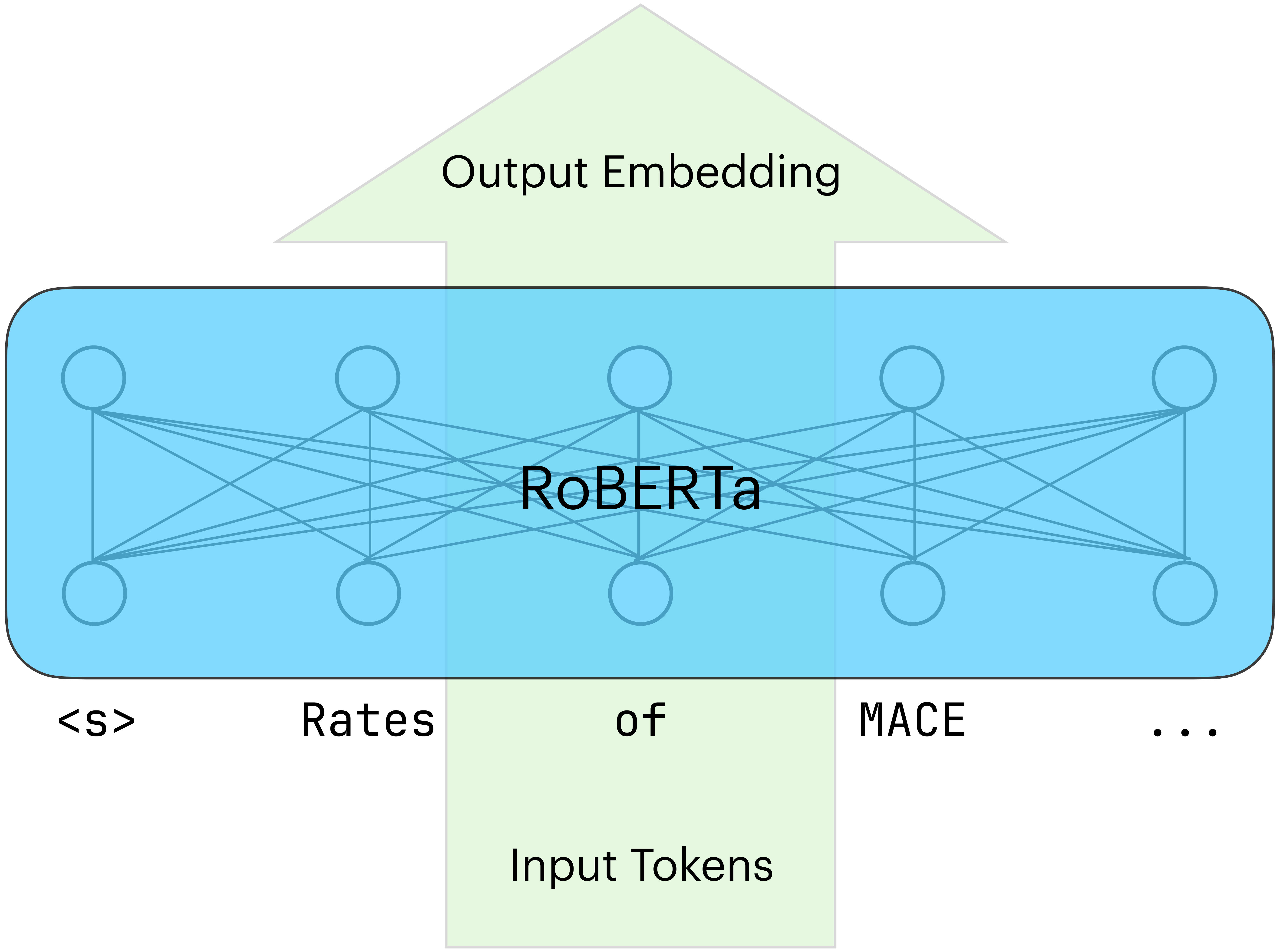
Methods

- NLP applications usually involve a pre-training step over a large amount of unlabeled data, then a fine-tuning step for your desired task
- RoBERTa is a pre-trained language model that is state-of-the-art in many NLP tasks
- We fine-tune RoBERTa to perform our Entity-Relation Extraction task, using our own dataset

RoBERTa is pre-trained on 160GB of Wikipedia entries, News articles, Reddit links, and Internet Stories

Entity-Relation Extraction Model Architecture

Methods

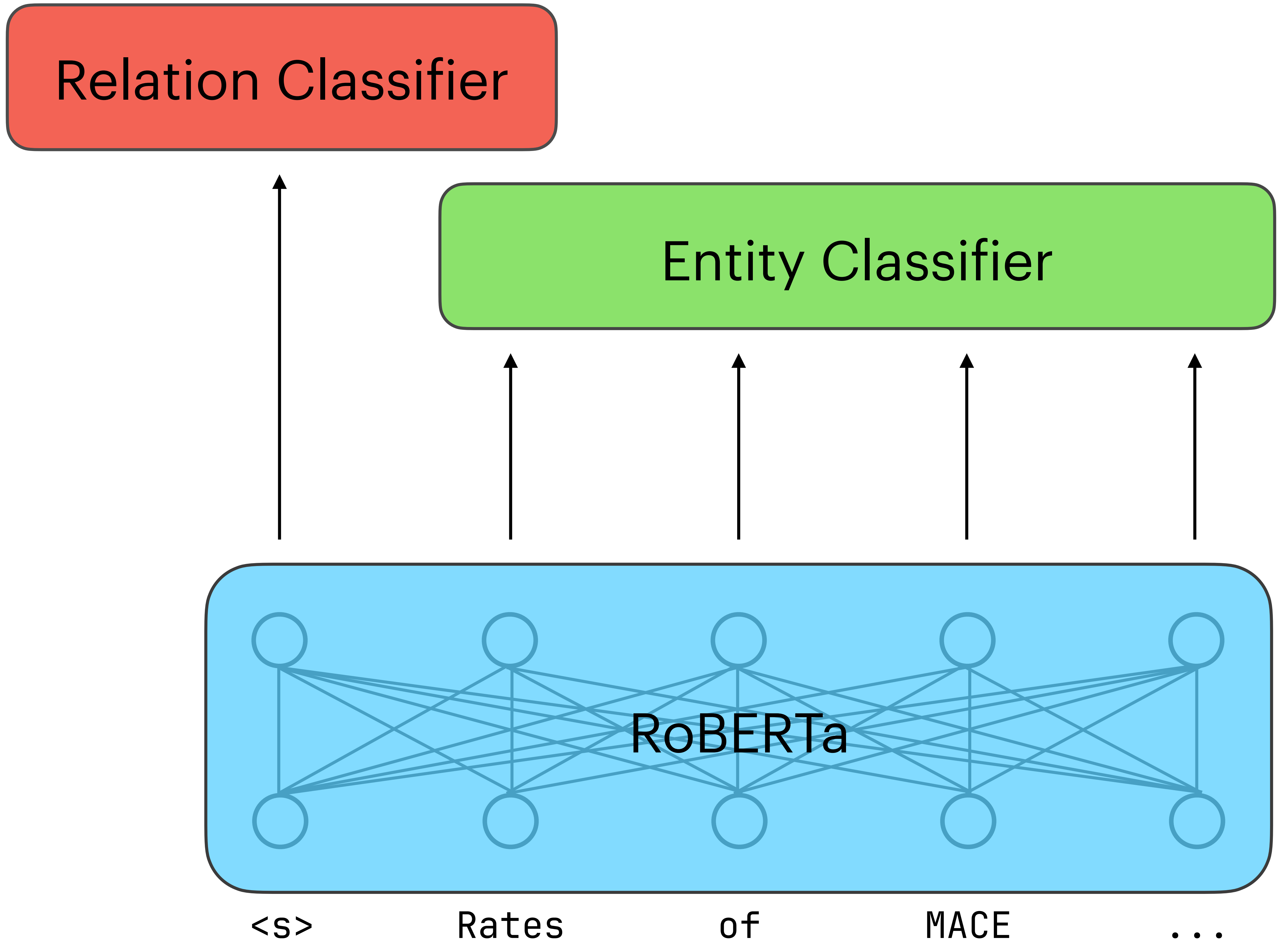


Entity-Relation Extraction Model Architecture

Methods

- **Entity Classifier:** Predict *Risk Estimates* and *Study Variables* for each input token
  - Uses Inside, Outside, Beginning (IOB2) labeling
- **Relation Classifier:** Predict the Relation (True or False) between a *Risk Estimate* and *Study Variable* pair
  - Query by inserting special tokens around entities, for example:

`Rates of $ MACE $ (RR ^ 0.91 ^ 95% ...`



Experimental Design

Methods

- Three RoBERTa variations: RoBERTa-Large (340M), RoBERTa-Base (110M), BioMed-RoBERTa-Base (110M)
- Two architectures tested for model training:
  - **Separate models:** one trained on **NER only,** the other trained on **RE only**
  - **Joint model:** model is trained on **both NER & RE** task (Joint Entity-Relation)
- Each model category was trained with 10 different seeds

- Top 3 models on validation set were then evaluated on held-out test set

Background

Objectives

Dataset

Methods

Results

Test Metrics

Results

- Now that we have trained a model, we can see how well it performs on our test dataset (annotated examples)

- **Named Entity Recognition:** Precision, Recall, F1-Score
- **Relation Extraction:** Matthew's Correlation Coefficient, Accuracy
- **Entity-Relation Extraction:** Jaccard Set Similarity

- In simpler terms, the closer the metric is to 1.0, the better

Test Metrics

Results

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 Score} = \frac{TP}{TP + 0.5\,(FP + FN)}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Matthew's Correlation Coefficient

$$= \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$\text{Jaccard Similarity} = \frac{|\text{Predicted} \cap \text{Actual}|}{|\text{Predicted} \cup \text{Actual}|}$$

Named Entity Recognition (NER) Task

Results

Model	Training Method	Precision	Recall	F1 Score
biomed-roberta-base	Separate	**0.774**	**0.789**	**0.781**
	Joint Entity Relation	0.745	0.789	0.767
roberta-base	Separate	**0.720**	**0.775**	**0.746**
	Joint Entity Relation	0.665	0.734	0.698
roberta-large	Separate	0.771	0.818	0.793
	Joint Entity Relation	**0.778**	**0.838**	**0.807**

Relation Extraction (RE) Task

Results

Model	Training Method	MCC	Accuracy
biomed-roberta-base	Separate	**0.929**	**0.967**
	Joint Entity Relation	0.927	0.967
roberta-base	Separate	**0.926**	**0.966**
	Joint Entity Relation	0.834	0.924
roberta-large	Separate	0.929	0.967
	Joint Entity Relation	**0.938**	**0.972**

Entity-Relation Extraction Task

Results

Model	Training Method	Similarity
biomed-roberta-base	Separate	**0.412**
	Joint Entity Relation	0.379
roberta-base	Separate	**0.340**
	Joint Entity Relation	0.265
roberta-large	Separate	0.433
	Joint Entity Relation	**0.443**

Discussion

Results

- RoBERTa-Large single model was the best for all tasks, achieving **0.81** F1 Score on NER, **0.93** MCC on RE, and **0.44** Jaccard Similarity on Entity-Relation Extraction
  - Impressive performance from just 95 training examples (<100KB of data)
- BioMed-RoBERTa-Base with separate models (NER-only + RE-only) had slightly worse performance than RoBERTa-Large single model (JER)
- Both BioMed-RoBERTa-Base and RoBERTa-Base two model (NER-only + RE-only) pipelines performed much better than their respective single model (JER) pipelines

Applications

- Researchers can semi-automate their literature reviews and meta-analysis to keep up with the increasing number of published articles
- This research will provide a framework for other Entity-Relation Extraction applications in medical fields (code will be open-sourced)
- Sets the foundation for a clinical knowledge database
  - Let users instantly aggregate, appraise, and summarize medical articles
- Next steps involve extracting study population, methodologies, and performing quality appraisal

Thank You for Listening! Any Questions?