Reconstruction of Coordination Ellipsis from Clinical Trial Eligibility Criteria Text

Using Natural Language Processing with Deep Learning

Haibo Yu (hy2628)[1], Zijian Wang(zw2606)[1], Yaotian Dai(yd2512)[1],
Tianyao Han(th2830)[1], and Jianqiong Zhan(jz3030)[1]
Mentors: Chunhua Weng[2] and Hao Liu[2]

1 Data Science Institute
COLUMBIA UNIVERSITY

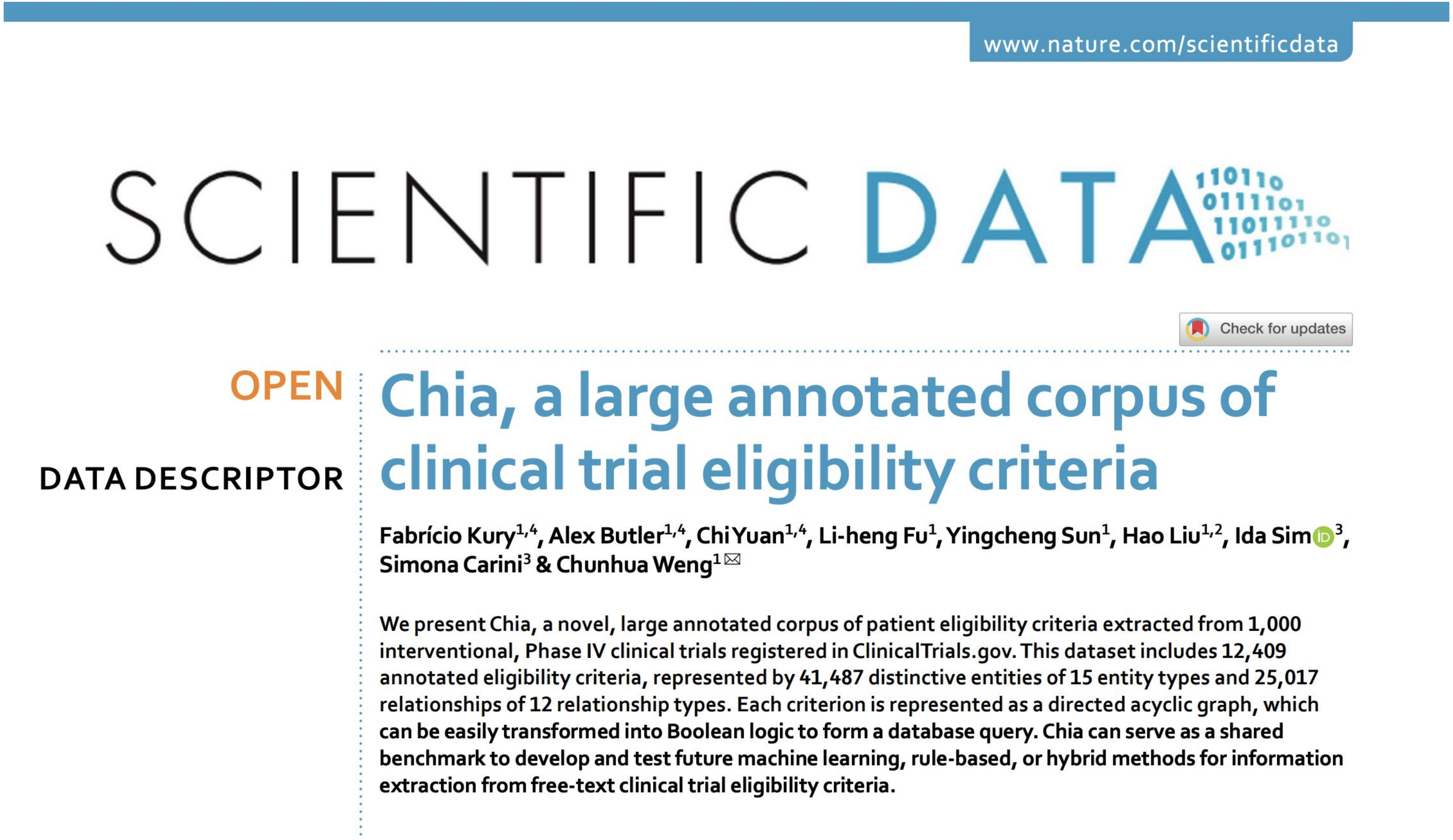
2 COLUMBIA UNIVERSITY
MEDICAL CENTER

NLP For Clinical Trial Summaries

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Example Coordination Ellipsis In Criteria Text

- Participant has active infection with **hepatitis B or hepatitis C virus**.
- **Histologically or cytologicallly confirmed breast cancer** that is metastatic of unresectable.
- Documented **germline mutation in BRA1 or BRCA2** that is predicted to be deleterious or suspected deleterious
- **major stomach or bowel resections**
- Participants with **a personal or family history of long QT syndrome**



www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

Chia, a large annotated corpus of clinical trial eligibility criteria

Fabrício Kury[1,4], Alex Butler[1,4], Chi Yuan[1,4], Li-heng Fu[1], Yingcheng Sun[1], Hao Liu[1,2], Ida Sim[3], Simona Carini[3] & Chunhua Weng[1] ✉

We present Chia, a novel, large annotated corpus of patient eligibility criteria extracted from 1,000 interventional, Phase IV clinical trials registered in ClinicalTrials.gov. This dataset includes 12,409 annotated eligibility criteria, represented by 41,487 distinctive entities of 15 entity types and 25,017 relationships of 12 relationship types. Each criterion is represented as a directed acyclic graph, which can be easily transformed into Boolean logic to form a database query. Chia can serve as a shared benchmark to develop and test future machine learning, rule-based, or hybrid methods for information extraction from free-text clinical trial eligibility criteria.

The Capstone Project Goals

- To leverage the CHIA dataset to improve biomedical NER
- To develop deep learning NLP models

Our Contributions

- We designed new tagging schemas for named entity recognition in Clinical Trial Summaries
    - We developed start-end labeling, E-label, Question Answering tagging schemas to mitigate ellipsis entity problems
    - We integrated domain knowledge from dictionary into model architecture

- We successfully extracted Ellipsis entities from CHIA

- Our enhanced model achieved 0.88 F1 score on CHIA data

Architecture

Model:

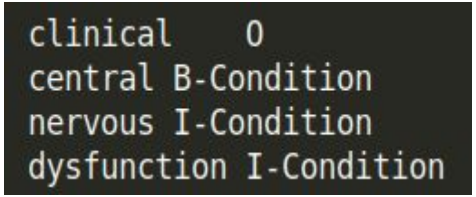
- None-Transformers(Flair, SpaCy,sklearn-crfsuite)
- Transformers(BERT, Bio-Clinical BERT, Pubmed-BERT)
- Ensemble models
- Dictionary enhancement

Data:

- CHIA, a human-annotated clinical trial corpus (the largest so far)
- COVID-19 for testing, the latest COVID-19 clinical trial corpus

Entity Labels:

- Standard labeling method:
  - B(Begin), I(Inside), O(Outside)
- Novel labeling method for ellipsis :
  - boundary label,
  - ellipsis label

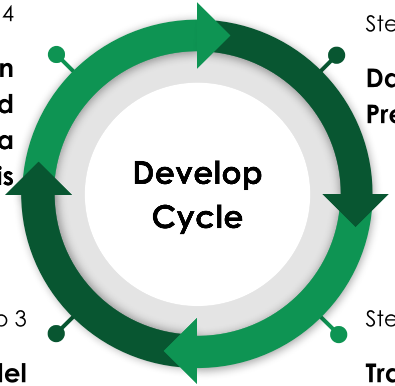


```
clinical    O
central B-Condition
nervous I-Condition
dysfunction I-Condition
```

Major activities

1.Online deployment
2.**UI: Demo**

Add datasets:
1.Dictionary
2.*Covid-19* dataset (generalizability)



Develop Cycle

Step 1
Data Preprocessing

1.Dictionary preprocessing
2.New labeling method
3.Multi-Domain preprocessing
4.Data Imbalance (refine)

Step 2
Training/Test Models

1.Training/Test on multi-domain entities
2.Covid-19 Test

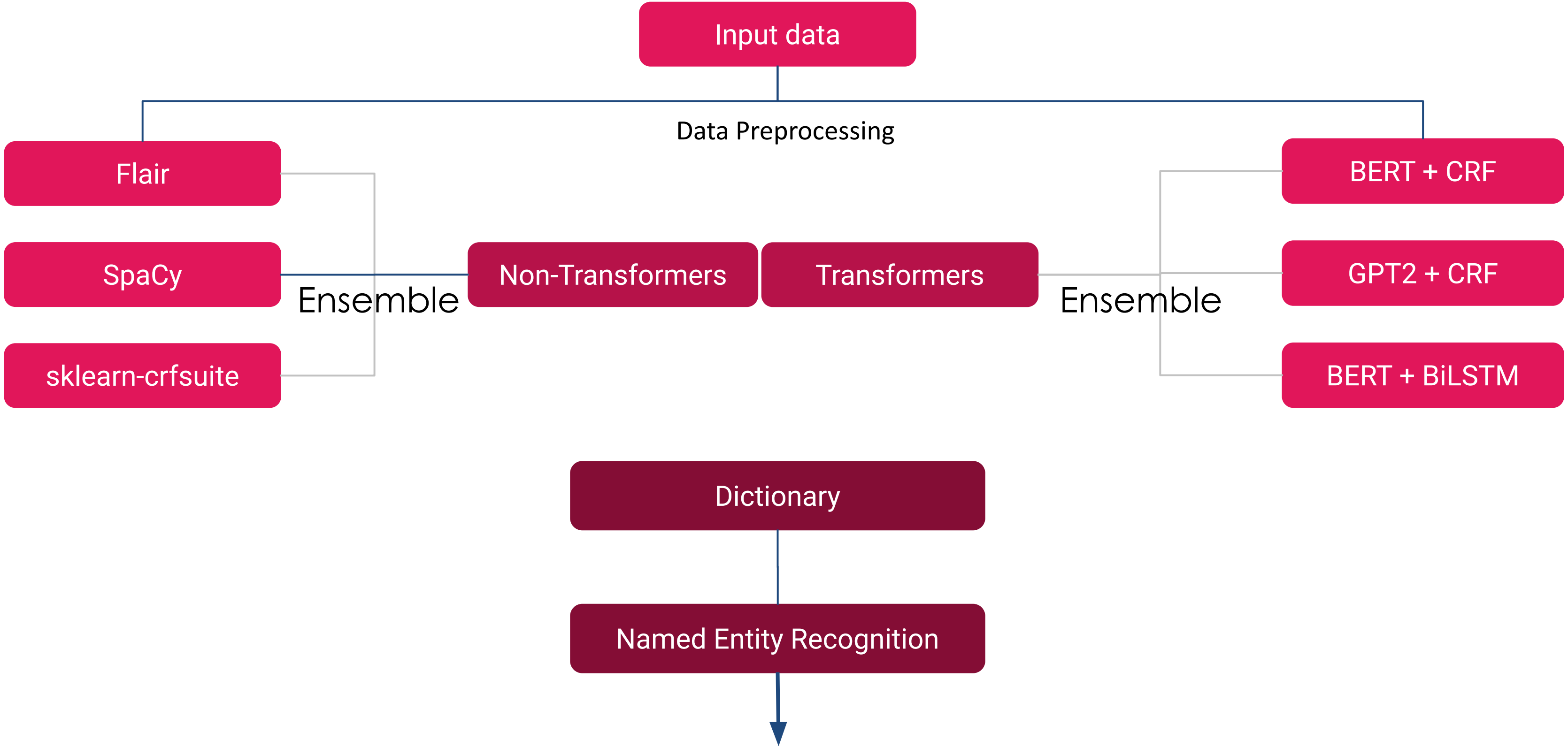
Step 3
Model Assembling

1.**Model Ensembles**
2.Non-BERT, BERT
3.Relaxed-ensemble:
  Token level
  Character level
4.vector-ensemble

Step 4
Evaluation and Data Analysis

1.Exact-match (new ver)
2.Relaxed-evaluation:
  Token level
  Character level
3.Definition: Inclusion Relation

Pipeline



Input data

Data Preprocessing

Flair

SpaCy

sklearn-crfsuite

Ensemble

Non-Transformers

Transformers

Ensemble

BERT + CRF

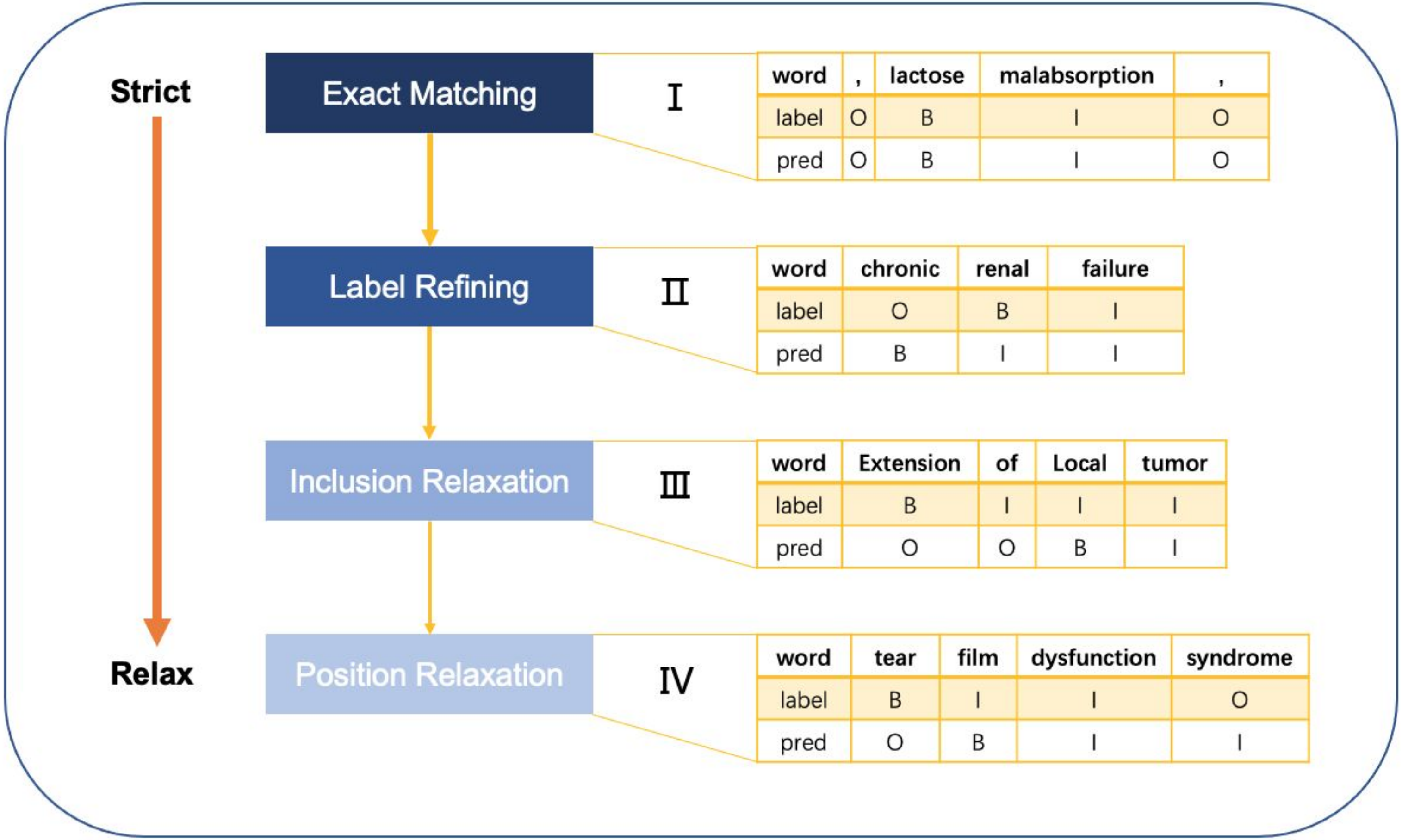
GPT2 + CRF

BERT + BiLSTM

Dictionary

Named Entity Recognition

Four Level Evaluation Architecture



Strict

Relax

Exact Matching

I

word	,	lactose	malabsorption	,
label	O	B	I	O
pred	O	B	I	O

Label Refining

II

word	chronic	renal	failure
label	O	B	I
pred	B	I	I

Inclusion Relaxation

III

word	Extension	of	Local	tumor
label	B	I	I	I
pred	O	O	B	I

Position Relaxation

IV

word	tear	film	dysfunction	syndrome
label	B	I	I	O
pred	O	B	I	I

Baseline

Datasets vs model performances

%	Flair	SpaCy	Microsoft-BERT	Flair+BERT	Flair+SpaCy
CHIA	79.42 | 85	74.98 | 83	78.73 | 87	77.33 | 86	77.45 | 80.52
COVID-19	75.48 | 85.41	65.30 | 77.64	\	\	68.67
	sk-crfsuite	GPT2+CRF	BERT+CRF	BERT+CNN	BERT+BiLSTM
CHIA	69.65 | 76	67.50	**82.23 | 87**	80.45	72.67
COVID-19	60.07	\	76.73	77.89	\

Baseline Error Analysis

1. ***HBV, HCV and HIV infections***:

Patient with Hepatitis B Virus (HBV) , Hepatitis C Virus (HCV) and Human
O O B I I O B O O B I I O B O O B
Immunodeficiency Virus (HIV) infections
I I O B O I

2. ***allergic disease or allergic reactions***:

history of allergic disease or reactions likely to be exacerbated by any
O O B I O I O O O O O O
component of the vaccine
O O O O

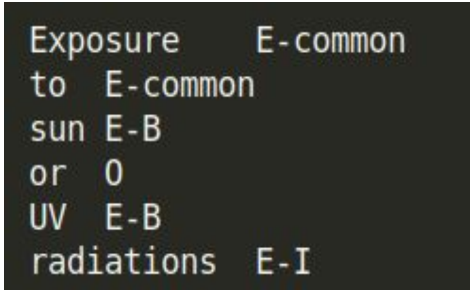
3. **alcohol or drug abuse**

Patients used to alcohol or drug (medication) abuse
O O O B O B O B O I

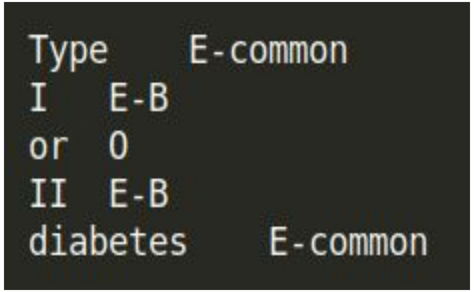
Enhanced Architecture - I

Model: BiLSTM + CRF

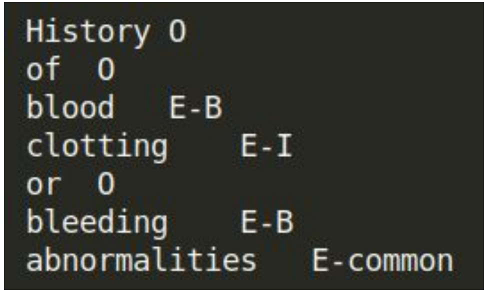
Label: Ellipsis labels (E-label)



```
Exposure    E-common
to  E-common
sun E-B
or  O
UV  E-B
radiations  E-I
```



```
Type    E-common
I   E-B
or  O
II  E-B
diabetes    E-common
```



```
History O
of  O
blood   E-B
clotting    E-I
or  O
bleeding    E-B
abnormalities   E-common
```

Enhanced - Ellipsis Result

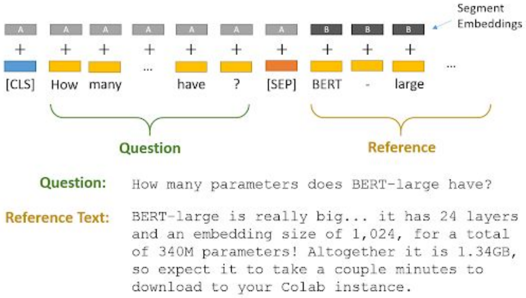
Datasets vs model performances Ellipsis Evaluation on CHIA

BiLSTM	entity	test sentences
Total	128	54
Exactly correct	96	41
Rate	0.75	0.76

Enhanced Architecture - II

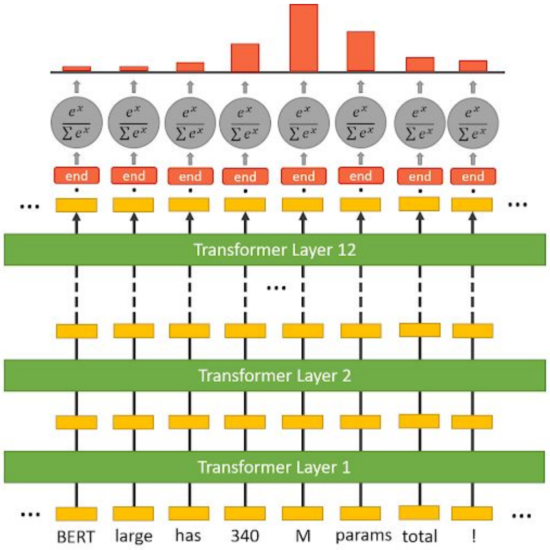
Model: Question Answering

Help extract boundary position.

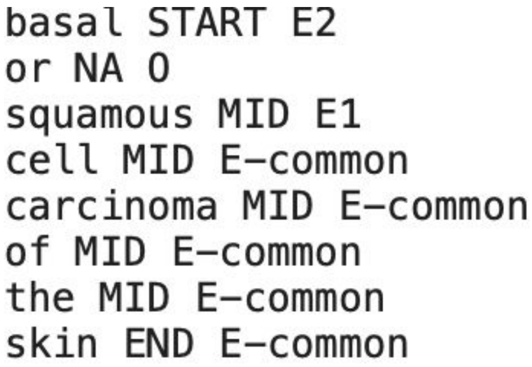


Question: How many parameters does BERT-large have?

Reference Text: BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.



Source: Chris McCormick



```
basal START E2
or NA O
squamous MID E1
cell MID E-common
carcinoma MID E-common
of MID E-common
the MID E-common
skin END E-common
```

Corresponding labeling method

Enhanced Architecture - II

Procedure to recover ellipsis entity:

1. Recognize the bounder

			start			end
Presence	of	serious	cardiac	or	respiratory	disease

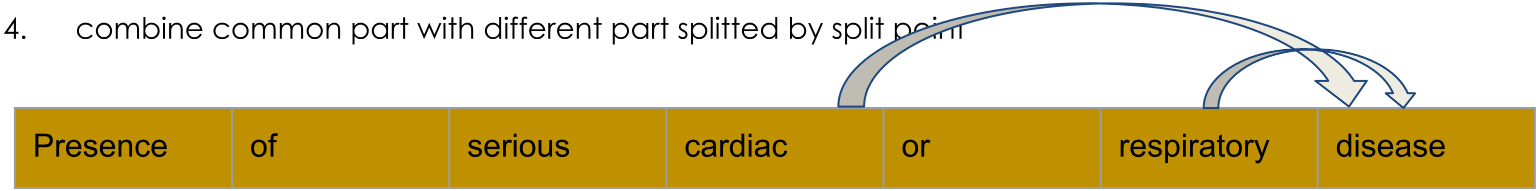
2. Finding split point like ‘OR’ ‘AND’ or punctuation

0	0	0	0	1	0	0

3. Finding common part

0	0	0	0	0	0	1

4. combine common part with different part splitted by split point



Presence	of	serious	cardiac	or	respiratory	disease

Ellipsis Test Result

Two type of errors:

1. boundary detection error:

Boundary detection error	boundary to be detected	correctly identified	rate
Total	54	48	88%

2. Ellipsis recovery error:

Ellipsis extraction error	entity to be extracted	correctly identified	rate
Total	108	97	89.8%

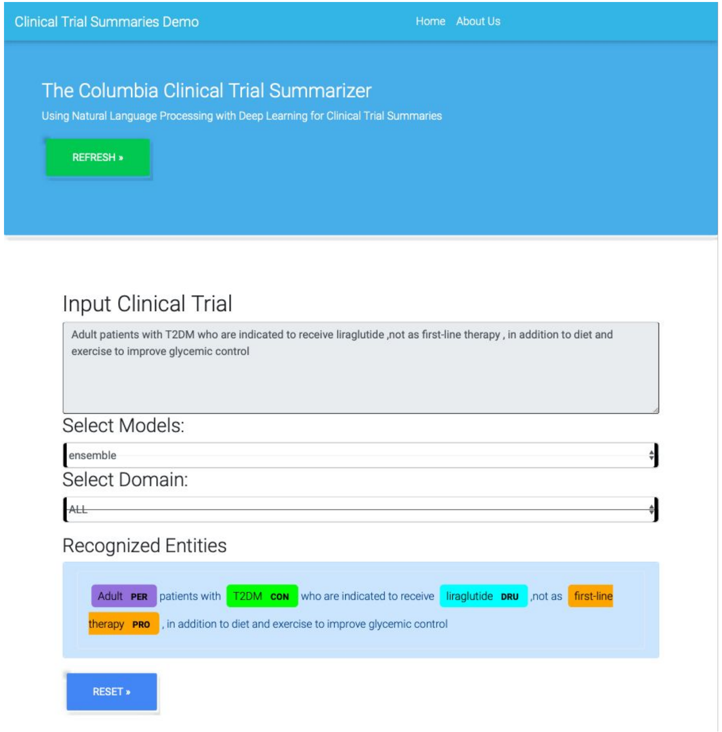
Discussion-Error Analysis(ellipsis)

1. Problems:
   a. “abnormal physical examination , vital signs or 12 lead ECG” (OOV)
   b. “Severe and/or chronic renal failure” (“and”, “/”)

2. In a nutshell, small training dataset

UI

A basic functioning version of the named entity tagging web application has been completed, which can be accessed at: http://34.121.40.143:9006. One can use the website to submit clinical trials to go through named entity recognition.



The Columbia Clinical Trial Summarizer

Using Natural Language Processing with Deep Learning for Clinical Trial Summaries

Input Clinical Trial

Adult patients with T2DM who are indicated to receive liraglutide ,not as first-line therapy , in addition to diet and exercise to improve glycemic control

Select Models: ensemble

Select Domain: ALL

Recognized Entities

Adult PER patients with T2DM CON who are indicated to receive liraglutide DRU ,not as first-line therapy PRO , in addition to diet and exercise to improve glycemic control

Summary

1. We have developed both traditional and deep learning models to the Named Entity Recognition task on CHIA data, in an End-to-End way.

2. We have implemented Question Answering, Dictionary, and Ensemble Learning to improve the boundary recognition

3. We have designed and implemented new labeling methods to solve Ellipsis Entity problems

4. We have proposed and implemented new evaluating methods for relaxed match

Contact us

Haibo Yu: hy2628@columbia.edu

Zijian Wang: zw2606@columbia.edu

Yaotian Dai: yd2512@columbia.edu

Tianyao Han: th2830@columbia.edu

Jianqiong Zhan: jz3030@columbia.edu

Thank You!