Natural Language 2
Structured Query Language

Umar Salman

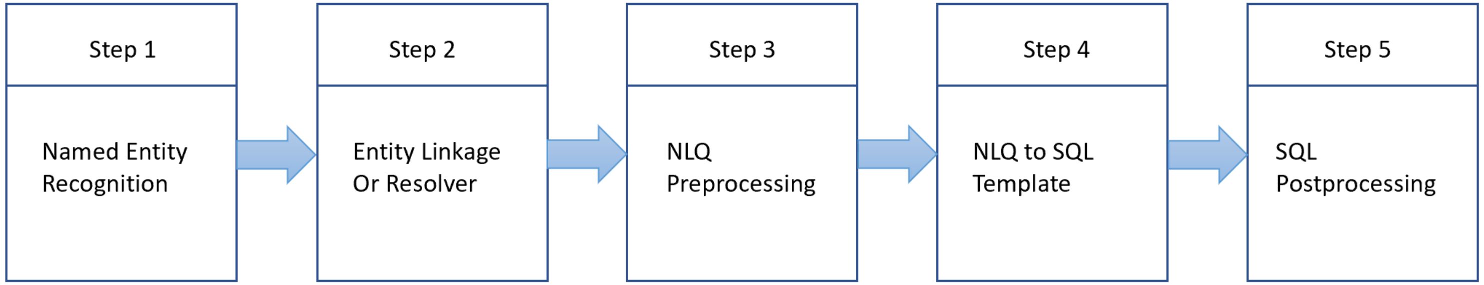
Agenda

- Goal
- Pipeline Architecture
- Datasets
  - Chia
  - Nostos
- CHIA
- Named Entity Recognition (NER)
- NOSTOS
- SQL Generation (T5)

Goal

- The goal of the project was to work towards the development of a natural language interface that can parse a user question or statement, transform it into a structured criteria representation and produce an executable clinical data query represented as an SQL query conforming to an EHR Common Data Model.

Architecture



1. Number of patients taking Aspirin
2. Aspirin -> Code (ICD9, ICD10, SnowMed)
3. Number of patients taking <ARG-DRUG><0>
4.
```
'SELECT COUNT( DISTINCT pe1.person_id) FROM (<SCHEMA>.person pe1 JOIN
(<DRUG-TEMPLATE><ARG-DRUG><0> JOIN <SCHEMA>.drug_exposure dr1 ON conc
ept_id=drug_concept_id) ON pe1.person_id=dr1.person_id);'
```
5.
```
"SELECT COUNT( DISTINCT pe1.person_id) FROM (cmsdesynpuf23m.person pe
1 JOIN (( SELECT descendant_concept_id AS concept_id FROM (SELECT * F
ROM (SELECT concept_id_2 FROM ( (SELECT concept_id FROM  cmsdesynpuf2
3m.concept WHERE vocabulary_id='RxNorm' AND ( concept_code='1191' ))
JOIN  ( SELECT concept_id_1, concept_id_2 FROM  cmsdesynpuf23m.concep
t_relationship WHERE relationship_id='Maps to' )  ON concept_id=conce
pt_id_1) ) JOIN cmsdesynpuf23m.concept ON concept_id_2=concept_id) JO
IN cmsdesynpuf23m.concept_ancestor ON concept_id=ancestor_concept_id
)  JOIN cmsdesynpuf23m.drug_exposure dr1 ON concept_id=drug_concept_i
d) ON pe1.person_id=dr1.person_id);"
```

Datasets

- CHIA
  - Annotated corpus of patient eligibility criteria extracted from 1,000 clinical trials
  - 41487 distinctive entities
  - 15 unique entity types
  - The entity categories are aligned with the domain names defined by the Observational Health Data Sciences and Informatics (ODHSI) OMOP CDM

Reference: https://www.nature.com/articles/s41597-020-00620-0

Datasets

- NOSTOS (Navigate OMOP-structured data via text-to-SQL)
  - The data consists of user generated questions and the corresponding SQL templates
  - The user generated questions are folded such that each sentence has synonyms words/phrases in them
  - There are 56 unique SQL queries which the user generated questions are trained on.

Reference: https://github.com/OHDSI/Nostos/tree/main/data

CHIA (NER)

- 1000 ann files
- 1000 text files

Pros

- Drug and Condition accurately represented
- Has relations between entities

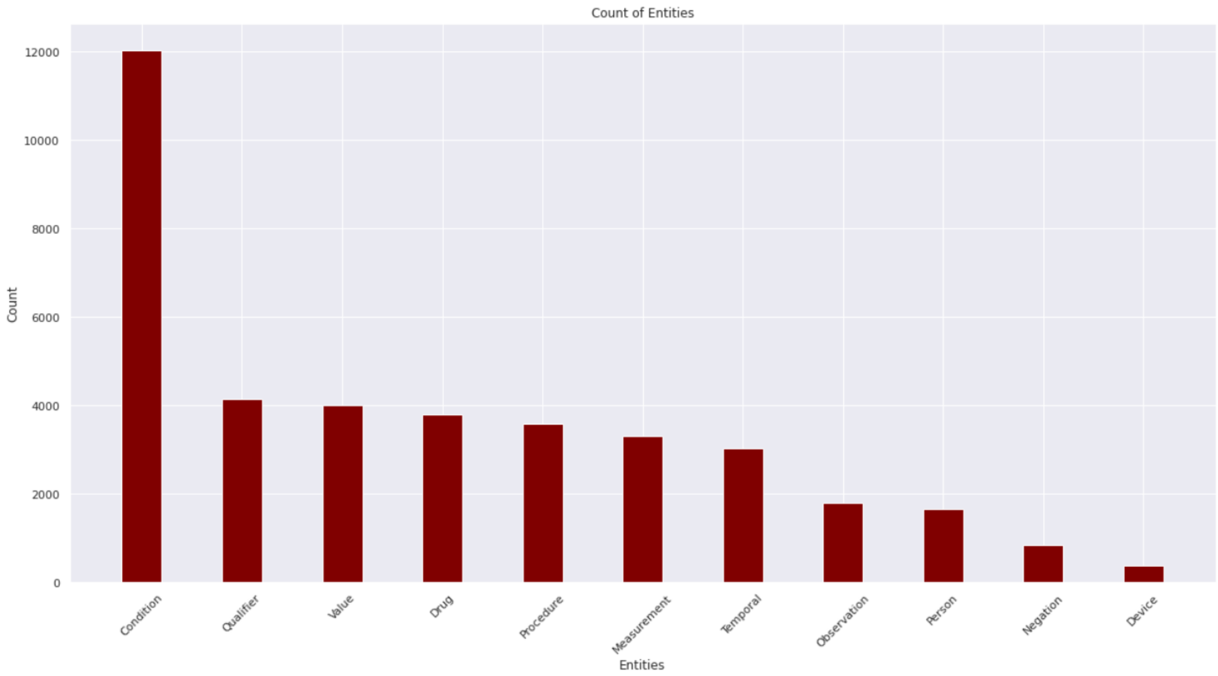
Cons

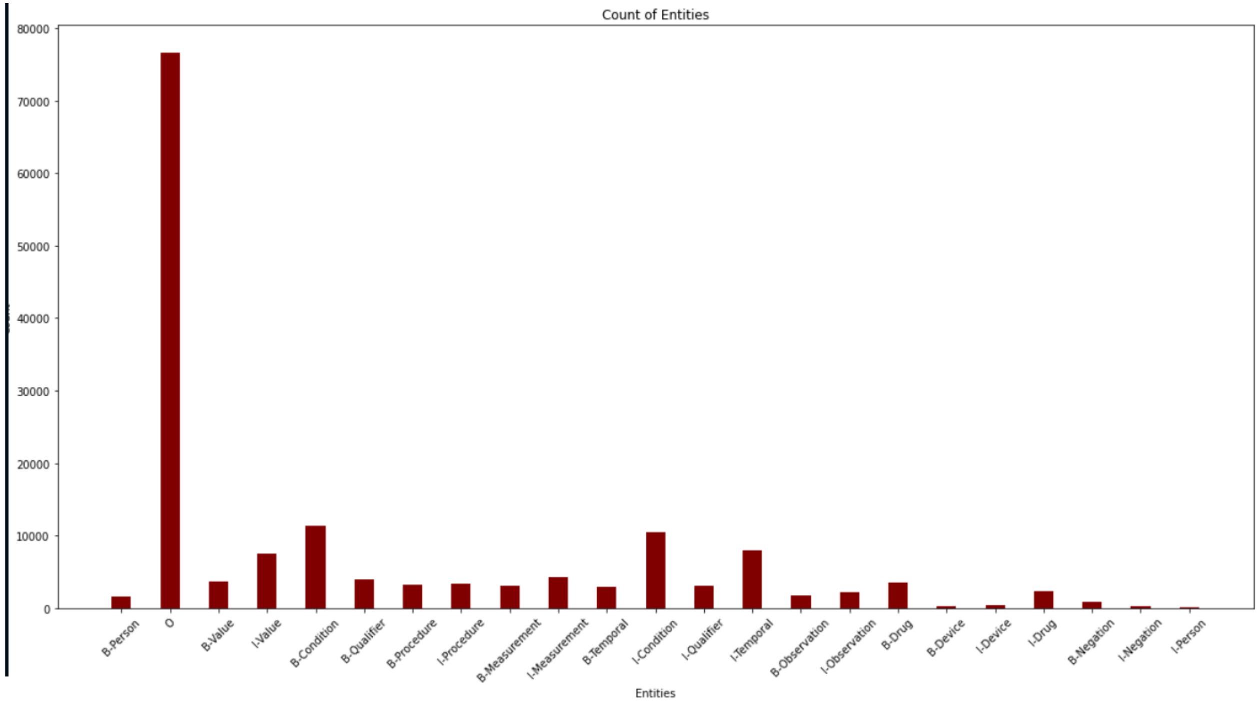
- Trouble with offsets leading to wrong labels
- Labelled data against each entity wasn't accurate (E.g. Temporal)
- Imbalanced data

```
T1	Condition 26 40 CNS metastases
T2	Condition 44 70 leptomeningeal involvement
*	OR T1 T2
T5	Procedure 144 151	treated
T6	Qualifier 164 179	been stable for
T7	Temporal 180 220	at least six months prior to study start
R1	Has_temporal Arg1:T6 Arg2:T7
*	OR T5 T6
T4	Condition 92 108	brain metastases
T8	Observation 238 248	history of
T9	Condition 249 265	brain metastases
R3	AND Arg1:T8 Arg2:T9
T10	Procedure 278 299	head CT with contrast
T11	Scope 238 265	history of brain metastases
A1	Optional T11
R4	AND Arg1:T11 Arg2:T10
T12	Condition 359 374	bone metastases
T13	Procedure 432 464	hepatic artery chemoembolization
T14	Temporal 465 489	within the last 6 months
T15	Temporal 491 500	one month
T16	Observation 514 547	other sites of measurable disease
R5	Has_context Arg1:T15 Arg2:T16
*	OR T14 T15
```

```
Patients with symptomatic CNS metastases or leptomeningeal involvement
Patients with known brain metastases, unless these metastases have been treated and/or have
been stable for at least six months prior to study start. Subjects with a history of brain
metastases must have a head CT with contrast to document either response or progression.
Patients with bone metastases as the only site(s) of measurable disease
Patients with hepatic artery chemoembolization within the last 6 months (one month if there
are other sites of measurable disease)
Patients who have been previously treated with radioactive directed therapies
Patients who have been previously treated with epothilone
Patients with any peripheral neuropathy or unresolved diarrhea greater than Grade 1
Patients with severe cardiac insufficiency patients taking Coumadin or other warfarin-
containing agents with the exception of low dose warfarin (1 mg or less) for the maintenance
of in-dwelling lines or ports
Patients taking any experimental therapies history of another malignancy within 5 years prior
to study entry except curatively treated non-melanoma skin cancer, prostate cancer, or
cervical cancer in situ
```

CHIA (NER)





Imbalance in entities

CHIA (NER)

File	Criteria	Text	Group_Entities	Relations	Tokens	Entities
NCT02186782_inc	inc	['Infertile women with eugonadotrophic anovula...	[('T2', 0, 9, 'Condition', 'Infertile'), ('T1'...	[('*', 'OR', 'T3 T4'), ('R1', 'Has_qualifier',...	[['Infertile', 'women', 'with', 'eugonadotroph...	([['Condition', 'Person', 'O', 'Qualifier', 'C...
NCT02186782_exc	exc	['Age < 20 or > 35 years.\n', 'Body mass index...	[('T1', 0, 3, 'Person', 'Age'), ('T2', 4, 8, '...	[('*', 'OR', 'T2 T3'), ('*', 'OR', 'T7 T8'), (...	[['Age', '<', '20', 'or', '>', '35', 'years'],...	([['Person', 'B-Value', 'I-Value', 'O', 'B-Val...
NCT02046395_inc	inc	['Type 2 Diabetes\n', 'Hypertension\n', 'Estim...	[('T1', 0, 15, 'Condition', 'Type 2 Diabetes')...	[('R1', 'Has_value', 'T3 T4'), ('*', 'OR', 'T5...	[['Type', '2', 'Diabetes'], ['Hypertension'], ...	([['B-Condition', 'I-Condition', 'I-Condition'...
NCT02046395_exc	exc	['Pregnancy\n', 'Patients with chronic kidney ...	[('T1', 0, 9, 'Condition', 'Pregnancy'), ('T2'...	[('R1', 'Has_value', 'T3 T4'), ('*', 'OR', 'T7...	[['Pregnancy'], ['Patients', 'with', 'chronic'...	([['Condition'], ['O', 'O', 'B-Condition', 'I-...
NCT02781610_inc	inc	['Male or female =18 years of age at Visit 1\n...	[('T1', 0, 4, 'Person', 'Male'), ('T2', 8, 14,...	[('R1', 'Has_value', 'T3 T4'), ('R3', 'Has_tem...	[['Male', 'or', 'female', '=18', 'years', 'of'...	([['Person', 'O', 'Person', 'B-Value', 'I-Valu...
...	...	...	...	...	...	...
NCT02321839_exc	exc	['Total lesion area of >12 DA or >30.5 mm2\n',...	[('T1', 0, 17, 'Measurement', 'Total lesion ar...	[('*', 'OR', 'T2 T3'), ('R2', 'Has_value', 'T5...	[['Total', 'lesion', 'area', 'of', '>12', 'DA'...	([['B-Measurement', 'I-Measurement', 'I-Measur...

Data Preparation and Processing

- Double check the word in text and see if entities with offsets match
- Removed irrelevant punctuation and corrupt/empty files
- Clean list of tokens to be used as input to the NER model
- Converted entity labels to NER format labels B- and I-
- Created a clean csv file for further use

Entities

- Condition
- Drug
- Procedure
- Measurement
- Observation
- Person
- Device
- Value
- Temporal
- Qualifier
- Negation

Relations

- OR
- AND
- Has_qualifier
- Has_value
- Has_negation
- Has_temporal
- Has_context

CHIA (NER)

	Tags	Sentence
0	[B-Person, O, B-Value, I-Value, I-Value, I-Value]	[ages, of, 7, and, 75, years]
1	[O, B-Condition, O, O, B-Qualifier, B-Qualifie...	[marked, disability, owing, to, primary, gener...
2	[B-Measurement, I-Measurement, O, B-Value, I-V...	[disease, duration, of, at, least, 5, years]
3	[B-Temporal, B-Procedure, I-Procedure]	[previous, brain, surgery]
4	[B-Condition, I-Condition, B-Value, I-Value, I...	[cognitive, impairment, <, 120, points, on, th...

```
Total Sentences: 12556
(Multiple sentences in a eligibility criteria file)

Train Test Split
Train: 8789
Validation: 1884
Test: 1883
```

Named Entity Recognition

Frameworks: PyTorch, Hugging Face

Models and Tokenizer: (Pretrained provided by HF)

- Bert (Baseline)
- BioBert
- MedBert
- SciBert
- BioBert (Large)

Architectural modifications and Finetuning Strategies

* Freezing/Unfreezing Pretrained Embeddings
* Adding extra layers after BERT/BioBert
* General labels Drug and Condition compared to B-Drug I-Drug (24 -> 13) (Including PAD and O)

Metrics:

Accuracy

F1 – Score

Loss:

Cross Entropy

Optimizer:

AdamW (Weight decay for regularization.)

Named Entity Recognition

We used two different criterions for our Evaluation Metric

- Strict Criteria
- Relaxed Criteria

In the Strict Criteria we look at the exact match between the gold annotated entity and the predicted entity.

In the Relaxed Criteria the predicted entity only overlaps with the gold annotated entity.

- We compare our results with a similar study done in "Transformer-Based Named Entity Recognition for Parsing Clinical Trial Eligibility Criteria"

https://dl.acm.org/doi/pdf/10.1145/3459930.3469560

Named Entity Recognition

“Transformer-Based Named Entity Recognition for Parsing Clinical Trial Eligibility Criteria”

Model	Strict Criterion			Relaxed Criterion		
	Precision	Recall	F1	Precision	Recall	F1
BERT	0.6052	0.6653	0.6339	0.7646	0.8132	0.7882
BERT-MIMIC	0.5934	0.6749	0.6316	0.7559	0.8228	0.7879
BLUEBERT	0.6244	0.6634	0.6433	0.7819	0.8033	0.7925
ALBERT	0.6007	0.6488	0.6238	0.7715	0.8020	0.7864
ALBERT-MIMIC	**0.6329**	0.6475	0.6401	**0.7871**	0.7818	0.7845
RoBERTa	0.6312	0.6818	0.6556	0.7715	0.8155	0.7929
RoBERTa-MIMIC	0.6158	0.6766	0.6448	0.7711	0.8175	0.7936
RoBERTa-MIMIC-Trial	0.6209	**0.6993**	**0.6578**	0.7662	**0.8333**	**0.7984**
ELECTRA	0.5749	0.6498	0.6101	0.7369	0.8013	0.7678
ELECTRA-MIMIC	0.6086	0.6723	0.6389	0.7661	0.8149	0.7897
Att-BiLSTM-CRF	0.3586	0.3896	0.3735	0.7064	0.7344	0.7201

Table 3: Performance of the transformer-based models vs. the baseline Att-BiLSTM-CRF model on Chia.

The best results we achieved were with the BioBert model, where we didn’t freeze the layers and didn’t add any layers to the pre-trained model where the pre-trained model can be found here: https://huggingface.co/dmis-lab/biobert-v1.1

- On the strict criteria it gave us a validation accuracy and F1 score of 77. 25% and 0.69 respectively.
- On the relaxed criteria it gave us a validation accuracy and F1 score of 82% and 0.77 respectively.

Named Entity Recognition

Per Entity

Model	Strict Criterion			Relaxed Criterion			
	Precision	Recall	F1	Precision	Recall	F1	
Overall	0.6209	0.6993	0.6578	0.7662	0.8333	0.7984	
Condition	0.7324	0.7878	0.7591	0.8721	0.9144	0.8928	0.90
Device	0.4667	0.6829	0.5545	0.6167	0.8293	0.7073	
Drug	0.6949	0.7910	0.7398	0.8418	0.9100	0.8746	0.9
Measurement	0.6127	0.6893	0.6487	0.7937	0.8464	0.8192	0.88
Mood	0.2727	0.2449	0.2581	0.3636	0.3265	0.3441	
Observation	0.2933	0.2444	0.2667	0.4333	0.3556	0.3906	
Person	0.6914	0.8643	0.7683	0.7257	0.8929	0.8007	0.86
Pregnancy_considerations	0.0000	0.0000	0.0000	0.3784	0.4444	0.4088	
Procedure	0.5012	0.6375	0.5612	0.6560	0.8031	0.7222	
Temporal	0.4800	0.6316	0.5455	0.6514	0.8008	0.7184	
Value	0.7000	0.7278	0.7136	0.8324	0.8685	0.8500	

Table 5: Performance of the best performing model (i.e., RoBERTa-MIMIC-Trial) by entity types on Chia.

	precision	recall	f1-score	support
Condition	0.88	0.91	0.90	3966
Device	0.67	0.65	0.66	82
Drug	0.89	0.90	0.90	1480
Measurement	0.91	0.86	0.88	1077
Negation	0.67	0.77	0.71	138
O	0.80	0.76	0.78	3175
Observation	0.50	0.45	0.47	383
Person	0.94	0.80	0.86	281
Procedure	0.75	0.81	0.78	1036
Qualifier	0.70	0.65	0.68	982
Temporal	0.69	0.79	0.73	507
Value	0.86	0.89	0.88	779
accuracy			0.82	13886
macro avg	0.77	0.77	0.77	13886
weighted avg	0.82	0.82	0.82	13886

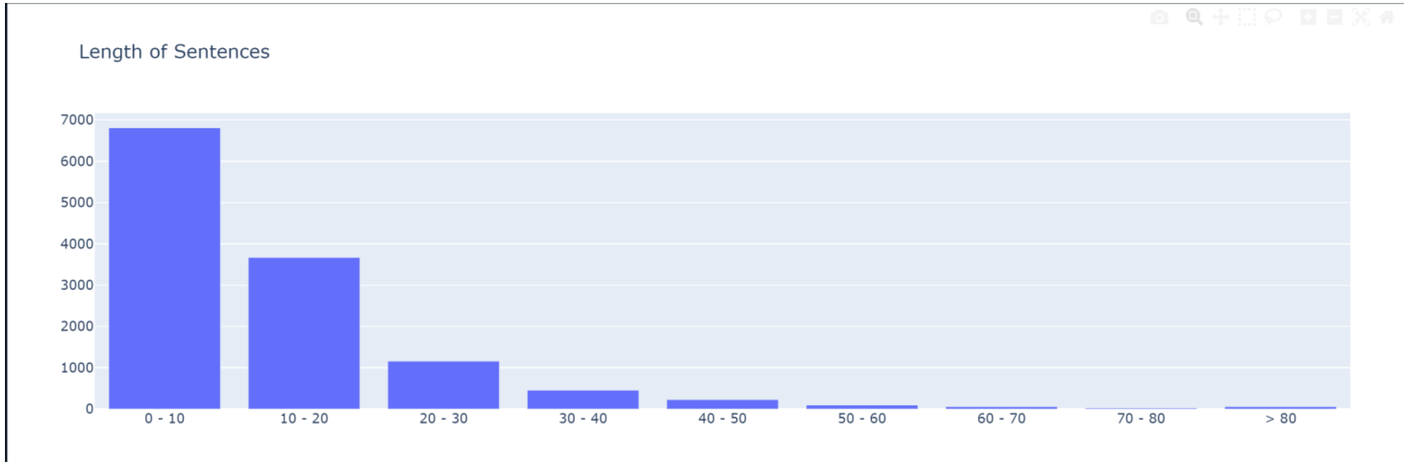
Named Entity Recognition

Comparison of the different models with different hyperparameters.

				Model & Parameters									Accuracy & Score			
Trial		Tokenizer	Model	Sentence Max Length	Batch Size	Epochs	Max Grad Norm	Learning Rate	Epsilon	LR Method	Optimizer		Training Loss	Validation Loss	Val Acc	Val F1 Score
Try 1		bert-base-cased	bert-base-cased	80	16	3	1	3.00E-05	1.00E-08	Linear Schedule with Warm-up	AdamW		0.115	0.175	76.20%	0.66
Try 2		bert-base-cased	bert-base-cased	80	16	5	1	3.00E-05	1.00E-08	Linear Schedule with Warm-up	AdamW		0.164	0.173	75.61%	0.65
Try 3		bert-base-cased	bert-base-cased	80	16	5	1	1.00E-05	1.00E-08	Linear Schedule with Warm-up	AdamW		0.151	0.185	74.67%	0.63
Try 4		bert-base-cased	bert-base-cased	80	16	10	1	3.00E-05	1.00E-08	Linear Schedule with Warm-up	AdamW		0.028	0.247	76.36%	0.675
Try 5		bert-base-cased	bert-base-cased	80	16	5	1	3.00E-04	1.00E-08	Linear Schedule with Warm-up	AdamW		0.95	1.18	0.00%	0.00%
Try 6		bert-base-cased	bert-base-cased	80	16	5	1	5.00E-05	1.00E-08	Linear Schedule with Warm-up	AdamW		0.055	0.196	76.63%	0.677
Try 7		bert-base-cased	bert-base-cased	80	8	5	1	5.00E-05	1.00E-08	Linear Schedule with Warm-up	AdamW		0.04	0.211	76.15%	0.678
Try 8		bert-base-cased	bert-base-cased	80	32	5	1	5.00E-05	1.00E-08	Linear Schedule with Warm-up	AdamW		0.74	0.188	76.57%	0.673
Try 9		fidukm34/biobert_v1.1_pubmed-f	dmis-lab/biobert-v1.1	80	16	5	1	3.00E-05	1.00E-08	Linear Schedule with Warm-up	AdamW		0.078	0.172	77.99%	0.7
Try 10		fidukm34/biobert_v1.1_pubmed-f	dmis-lab/biobert-v1.2	80	16	5	1	5.00E-05	1.00E-08	Linear Schedule with Warm-up	AdamW		0.054	0.186	77.60%	0.698
Try 11		fidukm34/biobert_v1.1_pubmed-f	dmis-lab/biobert-v1.3	80	16	3	1	3.00E-05	1.00E-08	Linear Schedule with Warm-up	AdamW		0.117	0.162	78.17%	0.699
Try 12		dmis-lab/biobert-v1.1	dmis-lab/biobert-v1.1	80	16	5	1	3.00E-05	1.00E-08	Linear Schedule with Warm-up	AdamW		0.07	0.176	77.64%	0.696
Try 13		dmis-lab/biobert-v1.1	dmis-lab/biobert-v1.1	80	16	5	1	3.00E-05	1.00E-08	None	AdamW	`	0.067	0.193	77.46%	0.698
Try 14		microsoft/BiomedNLP-PubMedBE	microsoft/BiomedNLP-	80	16	5	1	3.00E-05	1.00E-08	None	AdamW		0.06	0.137	76.20%	0.604
Try 15		sciarrilli/biobert-base-cased-v1.2-f	dmis-lab/biobert-v1.1	80	16	5	1	0.00005	0.00000001	None	AdamW	`	0.079	0.169	0.7811	0.693
Try 16		emilyalsentzer/Bio_ClinicalBERT	dmis-lab/biobert-v1.1	80	16	5	1	0.00005	0.00000001	None	AdamW		0.079	0.169	0.7811	0.693
Try 17		dmis-lab/biobert-v1.1	fidukm34/biobert_v1.1	80	16	5	1	0.00005	0.00000001	None	AdamW		0.084	0.175	0.7803	0.695
Try 18		sciarrilli/biobert-base-cased-v1.2-f	sciarrilli/biobert-base-c	80	16	5	1	0.00005	0.00000001	None	AdamW		0.071	0.184	0.7626	0.684
Try 19	Freeze/Add	fidukm34/biobert_v1.1_pubmed-f	dmis-lab/biobert-v1.1	80	16	20	1	0.00005	0.00000001	None	AdamW		0.259	0..23	68.32	0.487
Try 20	Freeze/Add	fidukm34/biobert_v1.1_pubmed-f	dmis-lab/biobert-v1.1	80	16	25	1	0.00003	0.00000001	None	AdamW		0.267	0.234	0.673	0.466
Try 21		dmis-lab/biobert-v1.1	dmis-lab/biobert-v1.1	80	16	10	1	0.003	0.00000001	Linear Schedule with Warm-up	SGD		0.15	0.169	0.7654	0.65
Try 22		dmis-lab/biobert-v1.1	dmis-lab/biobert-v1.1	80	16	5	1	0.00005	0.00000001	None	AdamW		0.06	0.19	0.7734	0.688
Try 23		dmis-lab/biobert-large-cased-v1.1	dmis-lab/biobert-large	80	16	5	1	0.00005	0.00000001	None	AdamW		0.14	0.255	0.7005	0.529

Named Entity Recognition

- The hyperparameters we selected were the following:
  - TOKENIZER_TYPE: dmis-lab/biobert-v1.1
  - MODEL_TYPE: dmis-lab/biobert-v1.1
  - MAX_LEN: 80
  - BATCH_SIZE: 16
  - EPOCHS: 5
  - MAX_GRAD_NORM: 1.0
  - LEARNING_RATE: 5e-05
  - EPSILON: 1e-08
  - TEST_SPLIT: 0.3
  - RANDOM_SEED: 42
  - OPTIMIZER: AdamW
  - LR_SCHEDULER: LinearWarmup
  - GENERAL_LABELS: False
  - ADDED_LAYERS: False
  - FREEZE_LAYES: False



Named Entity Recognition

```
NER PHASE
-----------------------
O        Count
O        of
O        patients
O        with
B-Drug   paracetamol
O        and
B-Drug   brufen
```

NOSTOS

Navigate OMOP-structured data via text-to-SQL

- The data consists of user generated questions and the corresponding SQL templates
- The user generated questions are folded such that each sentence has synonyms words/phrases in them
- There are 56 unique SQL queries which the user generated questions are trained on.

- Natural Language Question
  - How many <SYN-ARG-patients/people/persons/individuals/subjects> <SYN-ARG-taking/take/are treated with/are on/under/receive/have treatment with/took/were treated with/were treated by/were on/were under/had treatment with/had received> <ARG-DRUG><0> or <ARG-DRUG><1>?

- SQL Query Generation
  - SELECT COUNT(DISTINCT dr1.person_id) FROM (<SCHEMA>.drug_exposure dr1 JOIN (<DRUG-TEMPLATE><ARG-DRUG><0> UNION <DRUG-TEMPLATE><ARG-DRUG><1>) ON dr1.drug_concept_id=concept_id);

NOSTOS

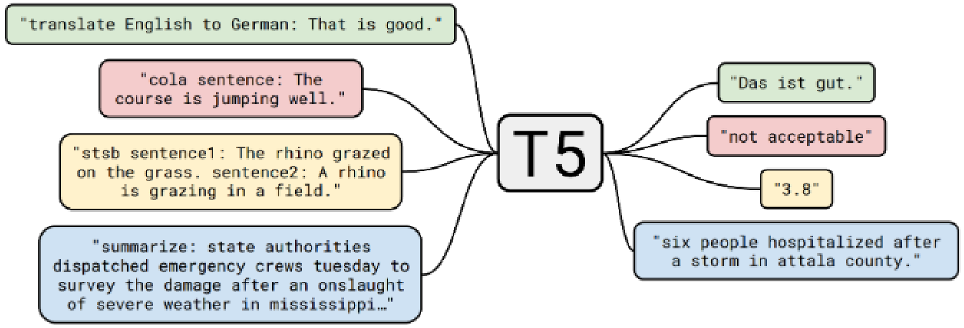
Sentence Combinations

```
Total Base Questions:
Train: 56, Val: 56, Test: 56

Total Folded Questions:
Train: 528, Val: 125, Test: 131

Total Query:
Train: 56, Val: 56, Test: 56
```

```
Train Length: 596961
Validation Length: 145368
Test Length: 56931
```

SQL Generation

- This uses a Seq2Seq model (Encoder + Decoder)
- T5 (Text-to-Text Transfer Transformer) - Transformer based architecture that uses a text-to-text approach
- T5 model that was pretrained with WikiSQL dataset



https://huggingface.co/mrm8488/t5-base-finetuned-wikiSQL

SQL Generation

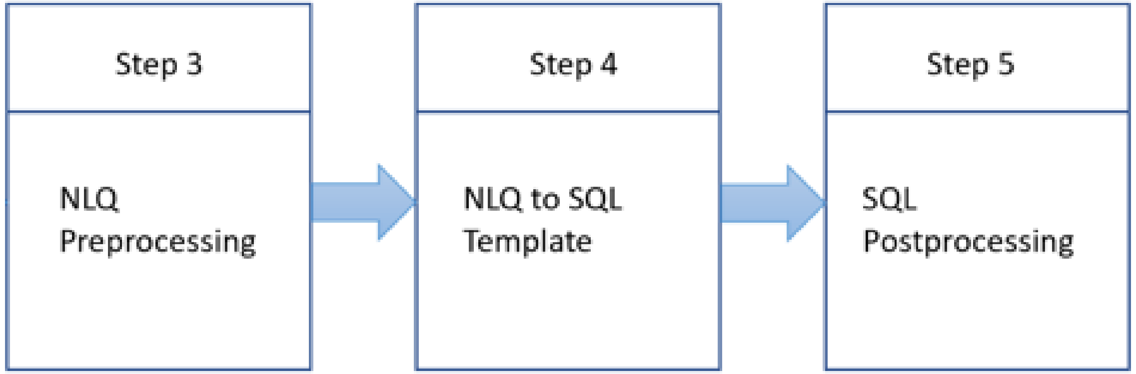
Pros

- Most likely will get an executable SQL Query if use template
- The NER and SQL Generation are both independent and can independently improve on both of them without them interfering with the other.

Cons

- If the SQL query predicts X Y Z entities and from the NER we get A B C or A Y Z or Y Z then the SQL query will fail to run

SQL Generation




```
'Number of patients taking <ARG-DRUG><0>'
```

```
'SELECT COUNT( DISTINCT pe1.person_id) FROM (<SCHEMA>.person pe1 JOIN
(<DRUG-TEMPLATE><ARG-DRUG><0> JOIN <SCHEMA>.drug_exposure dr1 ON conc
ept_id=drug_concept_id) ON pe1.person_id=dr1.person_id);'
```

```
"SELECT COUNT( DISTINCT pe1.person_id) FROM (cmsdesynpuf23m.person pe
1 JOIN (( SELECT descendant_concept_id AS concept_id FROM (SELECT * F
ROM (SELECT concept_id_2 FROM ( (SELECT concept_id FROM  cmsdesynpuf2
3m.concept WHERE vocabulary_id='RxNorm' AND ( concept_code='1191' ))
JOIN  ( SELECT concept_id_1, concept_id_2 FROM  cmsdesynpuf23m.concep
t_relationship WHERE relationship_id='Maps to' )  ON concept_id=conce
pt_id_1) ) JOIN cmsdesynpuf23m.concept ON concept_id_2=concept_id) JO
IN cmsdesynpuf23m.concept_ancestor ON concept_id=ancestor_concept_id
)  JOIN cmsdesynpuf23m.drug_exposure dr1 ON concept_id=drug_concept_i
d) ON pe1.person_id=dr1.person_id);"
```

SQL Generation

- Fine tuned the T5 model on the pre-trained model on the Nostos data
- Created custom datasets and dataloaders
- Added special tokens to our tokenizer (e.g. [ARG-DRUG] etc)

https://github.com/amazon-research/nl2sql-omop-cdm

SQL Generation

Model Hyperparameters

- MODEL_NAME: mrm8488/t5-base-finetuned-wikiSQL
- TOKENIZER_NAME: mrm8488/t5-base-finetuned-wikiSQL
- MAX_INPUT_LENGTH: 256
- MAX_OUTPUT_LENGTH: 512
- TRAIN_BATCH_SIZE: 8
- EVAL_BATCH_SIZE: 8
- EPOCHS: 5
- LEARNING_RATE: 0.001
- EPSILON: 1e-08
- RANDOM_SEED: 42
- WEIGHT_DECAY: 0.01
- MAX_GRAD_NORM: 1.0
- OPTIMIZER: AdamW
- LR_SCHEDULER: LinearWarmup
- FREEZE_ENCODER: False
- FREEZE_EMBEDDINGS: False

Mathes: 5427 Total Count: 5550
Exact Match Accuracy: 97.78%

SQL Generation

```
-----------------------
PREPROCESSING PHASE
-----------------------
paracetamol
        <ARG-DRUG><0>
brufen
        <ARG-DRUG><1>

Count of patients with <ARG-DRUG><0> and <ARG-DRUG><1>
-----------------------
SQL GENERATION PHASE
-----------------------
SELECT COUNT( DISTINCT dr1.person_id) FROM ((<SCHEMA>.drug_exposure dr1 JOIN <DRUG-TEMPLATE><ARG-DRUG><0> ON
```