Ontology and Research Medical Data

Poonam Sampat
poonam.sampat@microsoft.com

Agenda

- Business Use Case
- Approach / Thinking for the Solution
- Data Science Process
- Demo

Business Use Case

"Medicine is an art whose magic and creative ability have long been recognized as residing in the interpersonal aspects of patient-physician relationship."

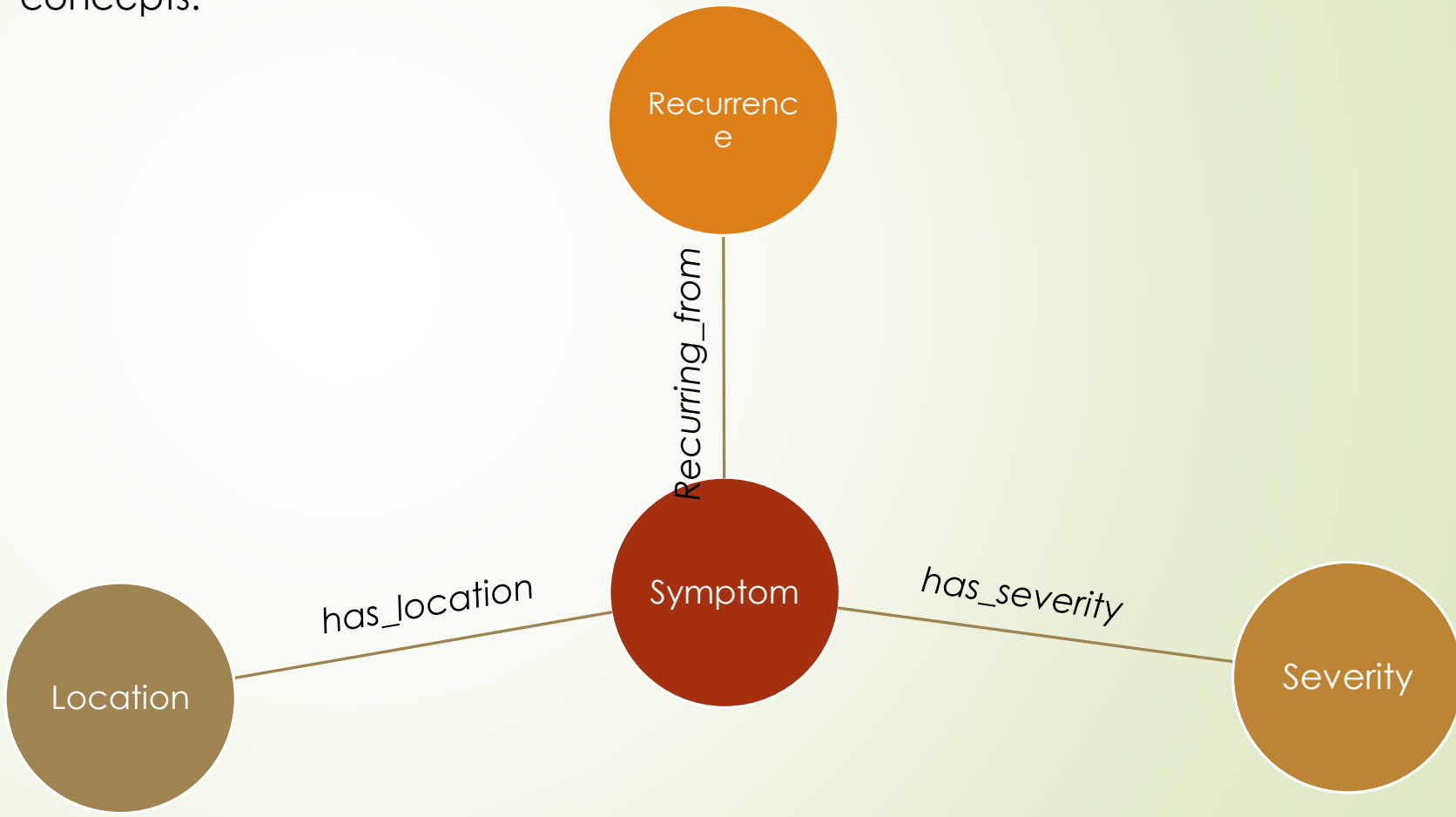
Mining the dialogue between a patient and doctor can help capture the essence of the communication and help building a knowledge base that can benefit the community at large.

4

Solution Approach

Ontology Based Knowledge Graph

Ontology is a data model that represents knowledge as a set of concepts within a domain and the relationships between these concepts.

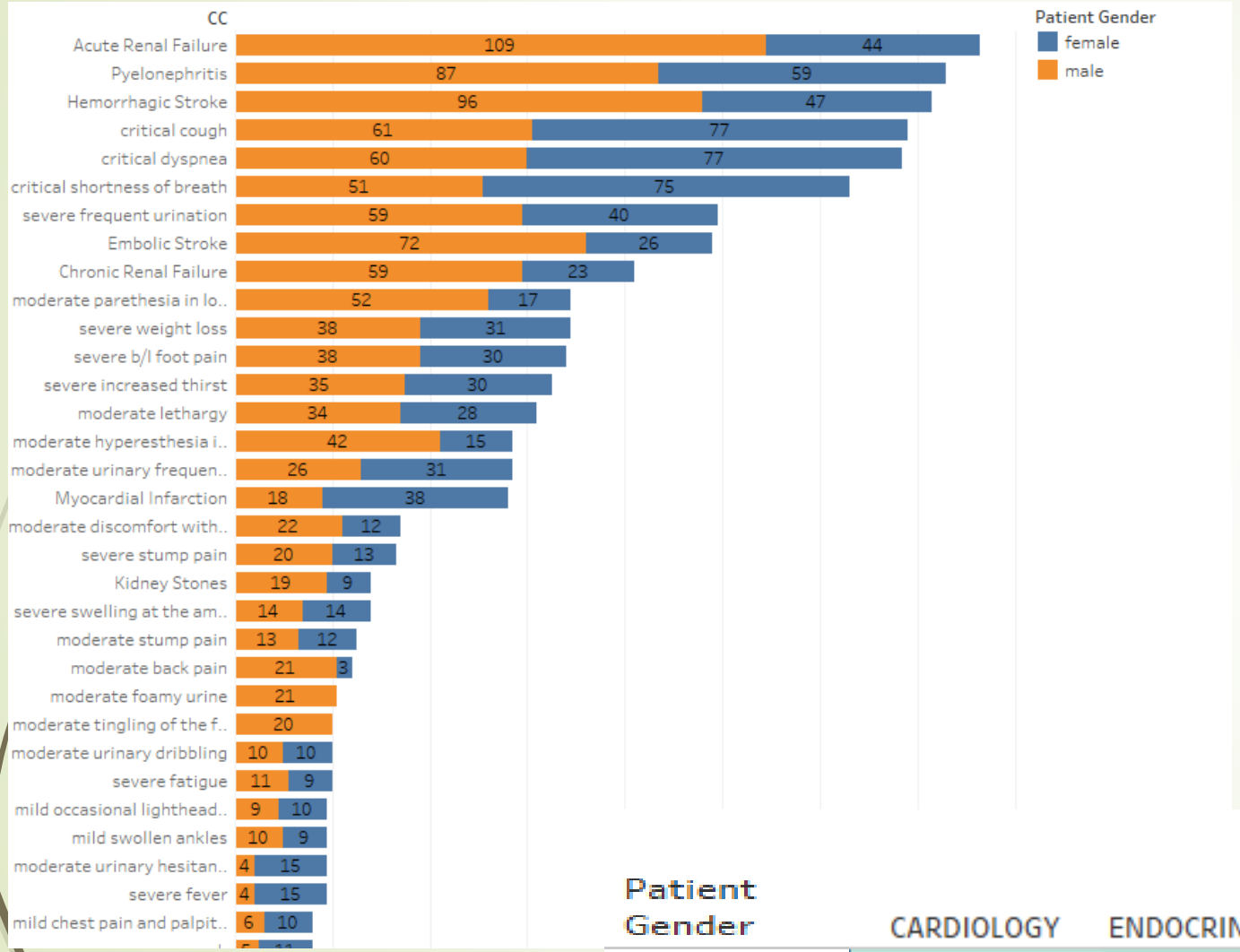


Data Science Process #1 Collect Data

- The medical data online had approximately 5500 records with 28 features
- The data maps symptoms and required lab tests to confirm diagnosed disease
- Field "SOAP_Note" in the dataset had relevant data
- Transcribed samples from https://www.mtsamples.com/

Example - SOAP_Note
s:a 33 year old female crystallographer presents with mild spells of vertigo, mild headaches particularly at the back of the head and in the morning x 2 weeks. pt also reports chronic mild occasional lightheadedness. o:Height 160 cm, Weight 53.8 kg, Temperature 37.3 C, Pulse 76, SystolicBP 146, DiastolicBP 93, Respiration 15, Heart = 2/6 systolic murmur at base of heart, Chest = clear to auscultation B/L, no rales or wheezing, Extremities = no edema or clubbing, Heart = normal S1, S2, RRR a:Hypertension p:performed E/M Level 2 (established patient) - Completed, and prescribed Hydrochlorothiazide - 50 mg po qd, and ordered Cholesterol.

Data Science Process - #2 Understand Data



Consult Ordered

Patient Gender	CARDIOLOGY	ENDOCRINOLOGY	NEPHROLOGY	PULMONARY DISEASE	THERAPY, PHYSICAL
female	44	83	16	86	8
male	86	119	48	55	9

3 Data Cleaning and Wrangling

- Create tokens out of the transcription provided.
- Remove stop words from the tokens created.
- Clean punctuations from the tokens(some punctuations are kept to maintain the sanity of sentence).
- Perform text extraction from data source and assign encounter ids which in turn help to classify the label of disease.
- Cleaning & Tokenization using NLTK.
- POS tagging

Algorithm Development – POS Tagging

- For Parts of speech tagging the data was divided into corpora into tokens.
- Then use Penn Tree bank to classify the tokens into one of the classes of Parts of Speech.
- This helps us to get form a semantic sentence out of the junk of words that are populated in the transcript.
- It also assists us in preparing for the Named Entity Recognition using a class of POS for example Diseases/Symptoms will always fall in the noun phrases and so on.
- We are using Syntax Net model for POS Tagging and dependency parsing and verifying manually for any errors.
- Chunking – IOB Format

	word	pos
3	33	CD
4	year	NN
5	old	JJ
6	female	JJ
7	crystallographer	NN
8	presents	NNS
9	with	IN
10	mild	JJ
11	spells	NNS
12	of	IN
13	vertigo	NN
14	,	,
15	mild	JJ
16	headaches	NNS

Algorithm Development – Defining Entities

- RECURRENCE (with values below)
  - Chronic
  - Occasional
  - Frequent
  - Recurrent
  - Increased frequency
  - x days
  - x weeks
  - x months
- months history
- weeks history
- days history
- for x days
- for x weeks
- for x months
- for x days

Algorithm Development – Defining Entities

SEVERITY (with values below)	SYMPTOM-LOC
• Critical • Severe • Moderate • Acute • Mild	• Foot • Lower limbs • In the ears • Of the feet • Back of the head

Algorithm Development – Defining Entities

- SYMPTOM
  - Headache
  - Lightheadedness
  - Cough
  - Fever
  - Dyspnea
  - Complaint
  - Paresthesia
  - Pyelonephritis
  - Pain
  - Ringing
  - Tingling
  - Routine Exam
  - Weight Loss
  - Renal Failure
  - Health Issues
  - Specific Complaints
  - Specific Issues
  - Foamy Urine
  - Frequent Urination
  - Increased thirst
  - Hemorrhagic Stroke
  - Chest pain
  - Spells of vertigo
  - Shortness of Breath
  - Type 2 Diabetes
  - Weak Urinary Stream
  - Changes to PMHPSH
  - Obstructive Pulmonary Disease

Chunking – IOB Format

The **IOB format** (short for inside, outside, beginning) is a common tagging format for tagging tokens in a chunking task in computational linguistics (ex. named-entity recognition).

Algorithm Development – Relationship Extraction

- Entity extraction is one part of the entire process. As soon as we have the entities annotated by our previous model, we are posed with another issue. Let us illustrate that with an example:
- … chronic headache at the back of the head and mild vertigo since **1 month**.
- … headache at the back of the head for **3 weeks** and mild vertigo since **1 month**.
- In the example above, the entities colored in red are Recurrence of the symptoms. In the absence of relation modelling, we won't be able to decide whether 3 weeks is the recurrence for headache or vertigo. The problem could be solved with using heuristics, but in the event of a new sentence formation, we will have to re-write our heuristic.

	word	pos tag	entity label	relation label
5	female	JJ	O	O
6	crystallographer	NN	O	O
7	presents	NNS	O	O
8	with	IN	O	O
9	mild	JJ	B-SEVERITY	a
10	spells	NNS	B-SYMPTOM	O
11	of	IN	I-SYMPTOM	O
12	vertigo	JJ	E-SYMPTOM	O
13	mild	JJ	B-SEVERITY	a
14	headaches	NNS	B-SYMPTOM	O
15	particularly	RB	O	O
16	at	IN	O	O

CRF - Conditional Random Field

Conditional Random Field is a model used to predict sequences.

They use contextual information from previous labels, thus increasing the amount of information the model has to make a good prediction.

Algorithm Development – Relationship Extraction

- So, we instead train another CRF that takes in inputs as Word, POS-tag, Entity Label and outputs Relation label which takes values ‘a’ and ‘b’ meaning above (Previous) and below (Next). Now, for each entity related to a symptom, we look at the relation and decide whether, say, severity is for the symptom immediately previous to it or next to the entity.

	word	pos tag	entity label	relation label
5	female	JJ	O	O
6	crystallographer	NN	O	O
7	presents	NNS	O	O
8	with	IN	O	O
9	mild	JJ	B-SEVERITY	a
10	spells	NNS	B-SYMPTOM	O
11	of	IN	I-SYMPTOM	O
12	vertigo	JJ	E-SYMPTOM	O
13	mild	JJ	B-SEVERITY	a
14	headaches	NNS	B-SYMPTOM	O
15	particularly	RB	O	O
16	at	IN	O	O

Algorithm Development – Creating the model

- Convert the words/tokens to a feature set by identify the pre and post token and identifying the POS tag for the same.
- Divide the data into training and test set with 80% of train set and the remaining as test set.
- Apply the feature set to CRF suite and use lbfgs algorithm with a threshold for max iterations as 100
- Apply a test feature to the model created and check for the accuracy using percentage of label match

```
{'bias': 1.0,
 'word.lower()': 's',
 'word[-3:]': 's',
 'word[-2:]': 's',
 'word.isupper()': False,
 'word.istitle()': False,
 'word.isdigit()': False,
 'postag': 'VBZ',
 'postag[:2]': 'VB',
 'BOS': True,
 '+1:word.lower()': 'a',
 '+1:word.istitle()': False,
 '+1:word.isupper()': False,
 '+1:postag': 'DT',
 '+1:postag[:2]': 'DT'}
```

Results CRF Model

- A white 42 year old lady complains about pain in the feet and occasional headaches that are mild on the back of the head. severe back pain too.

A	O
white	O
42	O
year	O
old	O
lady	O
complains	O
about	O
pain	B-SYMPTOM
in	B-SYMPTOM-LOC
the	I-SYMPTOM-LOC
feet	E-SYMPTOM-LOC
and	O
occasional	B-RECURRENCE
headaches	B-SYMPTOM
that	O
are	O
mild	B-SEVERITY
on	O
the	O
back	B-SYMPTOM-LOC
of	I-SYMPTOM-LOC
the	I-SYMPTOM-LOC
head	E-SYMPTOM-LOC
severe	B-SEVERITY
back	O
pain	B-SYMPTOM
too	O

Alternate Model – Cliner Model

- Clinical Named Entity Recognition system (CliNER) is an open-source natural language processing system for named entity recognition in clinical text of electronic health records. CliNER system is designed to follow best practices in clinical concept extraction, as established in i2b2 2010 shared task.
- CliNER is implemented as a sequence classification task, where every token is predicted IOB-style as either: Problem, Test, Treatment, or None. Command line flags let you specify two different sequence classification algorithms:
- 1. CRF (default) - with linguistic and domain-specific features
- 2. LSTM.

Alternate Model – Cliner Model Parameters

- load_all_pretrained_token_embeddings False
- load_only_pretrained_token_embeddings False
- tagging_format bio
- use_character_lstm True
- use_crf True
- Use_LSTM True
- use_features_before_final_lstm False
- character_embedding_dimension 25
- character_lstm_hidden_state_dimension 25
- token_embedding_dimension 100
- freeze_token_embeddings False
- token_lstm_hidden_state_dimension 100
- optimizer sgd
- gradient_clipping_value 5.0
- remap_unknown_tokens_to_unk True
- learning_rate 0.005

Results Cliner Model

- The patient is an elderly thin white female, very pleasant, in no acute distress. VITAL SIGNS: Her temperature is 98.8 and vital signs are all stable, within normal limits. HEENT: Head is grossly atraumatic and normocephalic. Sclerae are anicteric. The conjunctivae are non-injected. NECK: Supple. CHEST: Clear. HEART: Regular rate and rhythm. ABDOMEN: Generally nondistended and soft. She is focally tender in the left lower quadrant to deep palpation with a palpable fullness or mass and focally tender, but no rebound tenderness. There is no CVA or flank tenderness, although some very minimal left flank tenderness. PELVIC: Currently deferred, but has history of grade 4 urinary bladder prolapse. EXTREMITIES: Grossly and neurovascularly intact.

	word	NER
0	her temperature	test
1	vital signs	test
2	tender in the left lower quadrant	problem
3	a palpable fullness	problem
4	mass	problem
5	rebound tenderness.	problem
6	cva	problem
7	grade __num__ urinary bladder	problem

Accuracy CRF Model Entity

	precision	recall	f1-score	support
B-RECURRENCE	0.998	1.000	0.999	587
E-RECURRENCE	0.997	1.000	0.998	301
I-RECURRENCE	1.000	1.000	1.000	58
B-SEVERITY	1.000	0.999	0.999	809
B-SYMPTOM	1.000	0.993	0.996	1531
E-SYMPTOM	1.000	0.997	0.999	784
I-SYMPTOM	1.000	1.000	1.000	307
B-SYMPTOM-LOC	1.000	1.000	1.000	184
E-SYMPTOM-LOC	1.000	1.000	1.000	76
I-SYMPTOM-LOC	1.000	1.000	1.000	27
avg / total	1.000	0.997	0.998	4664

Accuracy CRF Model Relation

	precision	recall	f1-score	support
a	0.909	0.962	0.935	706
b	0.894	0.834	0.863	283
avg / total	0.905	0.925	0.914	989

Applications

- The product can be used as a doctor's assistant. As soon as a patient is there for a check-up the assistant will read the transcription histories and will help the doctor diagnose better based on the symptoms identified.
- The knowledge Base formed for the application can then further be used to create a chatbot which can help the patient/doctor ask a query which will be answered.

THANK YOU