



CogStack

 KING'S HEALTH PARTNERS

KING'S
College
LONDON



Medical Concept Annotation Toolkit

[\(\[github.com/CogStack/MedCAT\]\(https://github.com/CogStack/MedCAT\)\)](https://github.com/CogStack/MedCAT)

[\(\[github.com/CogStack/MedCATtrainer\]\(https://github.com/CogStack/MedCATtrainer\)\)](https://github.com/CogStack/MedCATtrainer)

[\(\[github.com/CogStack/MedCATservice\]\(https://github.com/CogStack/MedCATservice\)\)](https://github.com/CogStack/MedCATservice)

Tom Searle

tom.searle@kcl.ac.uk

Based off original slides by Zeljko Kraljevic (zeljko.kraljevic@kcl.ac.uk)

Motivation

- Query examples
 - All patients with Atrial Fibrillation (AF, Afib, A. Fibrillation, ...)
 - Includes spelling mistakes, acronyms, shorthand etc.
 - Contextualised mentions, i.e.:
 - Status = Affirmed
 - Experiencer = patient
 - Temporality = Current
- Use-cases
 - Clinical research questions
 - Phenotyping
 - Disease prevalence
 - Down stream aggregate analysis of symptom / finding / diagnosis / drug / procedure etc.
 - Clinical coding
 - Clinical trials - cohort selection / recruitment

Two Main Tasks

- Document Annotation
 - Entity Extraction (NER)
 - Entity Linking (+L)
- Concept Classification

Patient X

PAST MEDICAL HISTORY: The patient denies high blood pressure, diabetes, heart disease, lung disease, thyroid, kidney, or bladder dysfunctions. The patient stated that she quit smoking prior to her past childbirth and is currently not pregnant. The patient has had a C-section and also an appendectomy.

MEDICATIONS: Patient currently states she is taking:

1. Vicodin 500 mg two times a day.
2. Risperdal.
3. Zoloft.
4. Stool softeners.
5. Prenatal pills.

Two Main Tasks

- Document Annotation

- Entity Extraction (NER): defines start / end boundaries for a concept
- Entity Linking (+L): links the text span with a single concept sourced from a configured knowledge base

- Concept Classification

- Contextualises a recognised concept, i.e. Negation

Patient c0030705 X

PAST MEDICAL HISTORY c0262926 : The patient c0030705 denies

high blood pressure c0020538 , diabetes c0011849 , heart disease c0018799 ,

lung disease c0024115 , thyroid c0040132 , kidney c0022646 , or bladder dysfunctions c0232841 .

The patient stated c0683521 that she quit smoking c0085134 prior to c0332152

her past c1444637 childbirth c0005615 and is currently not pregnant c0549206 .

The patient has c0332310 had a C-section c0007876 and also an

appendectomy c0003611 .

MEDICATIONS c0013227 : Patient c0030705 currently c0521116 states c3148680 she is taking:

1. Vicodin c0483514 500 mg c0024467 two times a day c0439511 .

2. Risperdal c0592071 .

3. Zoloft c0284660 .

4. Stool c0015733 softeners.

5. Prenatal c0678804 pills c0994475 .

Two Main Tasks

- Document Annotation
 - Entity Extraction and Linking
- Concept Classification
 - Negation, ...

pulmonary disease



cui C0024115

tui T047

type Disease or Syndrome

source_value lung disease

acc 1

Negated True

Experiencer Patient

Patient C0030705 X

PAST MEDICAL HISTORY C0262926 : The patient C0030705 denies

high blood pressure C0020538 , diabetes C0011849 , heart disease C0018799 ,

lung disease C0024115 , thyroid C0040132 , kidney C0022646 , or bladder dysfunctions C0232841 .

The patient stated C0683521 that she quit smoking C0085134 prior to C0332152

her past C1444637 childbirth C0005615 and is currently not pregnant C0549206 .

The patient has C0332310 had a C-section C0007876 and also an

appendectomy C0003611 .

MEDICATIONS C0013227 : Patient C0030705 currently C0521116 states C3148680 she is taking:

1. Vicodin C0483514 500 mg C0024467 two times a day C0439511 .

2. Risperdal C0592071 .

3. Zolofl C0284660 .

4. Stool C0015733 softeners.

5. Prenatal C0678804 pills C0994475 .

MedCAT: Concept Annotation (Entity Extraction and Linking)

Concept Database		
<i>ID</i>	<i>Names</i>	<i>Embedding</i>
1	Heart Rate, HR	V_1
2	Hour, HR	V_2
...		

The procedure took 3 hours,
during which his heart rate was in
the 60s.

Later in the evening the HR was in
the 50s.

- .
- .
- .

MedCAT: Training Procedure

Concept Database		
<i>ID</i>	<i>Names</i>	<i>Embedding</i>
1	Heart Rate, HR	V_1
2	Hour, HR	V_2
...		

The procedure took 3 hours, during which his heart rate was in the 60s. Later in the evening the HR was in the 50s.

$$V_{cntx} = \frac{1}{2s} \left[\sum_{i=1}^s V_{w_k-i} + \sum_{i=1}^s V_{w_k+1+i} \right]$$

$$sim = \max(0, \frac{V_{concept}}{\|V_{concept}\|} \cdot \frac{V_{cntx}}{\|V_{cntx}\|})$$

$$lr = \frac{1}{C_{concept}}$$

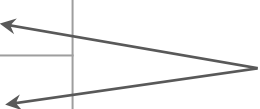
$$V_{concept} = V_{concept} + lr \cdot (1 - sim) \cdot V_{cntx}$$

MedCAT: Run Phase

Concept Database		
<i>ID</i>	<i>Names</i>	<i>Embedding</i>
1	Heart Rate, HR	V_1
2	Hour, HR	V_2
...		

The procedure took 3 hours,
during which his heart rate was in
the 60s.

Later in the evening the HR was in
the 50s.


$$V_{ctx} = \frac{1}{2s} \left[\sum_{i=1}^s V_{w_k-i} + \sum_{i=1}^s V_{w_k+1+i} \right]$$

Word Embeddings

<i>Disease -> Medication</i>	<i>Disease -> Procedure</i>	<i>Symptom -> Medication</i>	<i>Symptom -> Everything</i>
Hypertensive disease	Neoplastic Process (Cancer)	Fever	Hemorrhage
Metoprolol 50 MG	Chemotherapy	Levofloxacin	Intracranial Hemorrhages
Metoprolol 25 MG	Radiosurgery	Vancomycin	Cerebellar hemorrhage
Valsartan 320 MG	FOLFOX Regimen	Vancomycin 750 MG	Postoperative Hemorrhage
Nadolol 20 MG	Chemotherapy Regimen	Azithromycin	Retroperitoneal Hemorrhage
Atenolol 100 MG	Preoperative Therapy	Levofloxacin 750 MG	Amyloid angiopathy
Enalapril 10 MG	Anticancer therapy	Dexamethasone	Internal bleeding
Oral form diltiazem	Parotidectomy	Lorazepam	Hematoma, Subdural, Chronic
nimodipine 30 MG	Resection of ileum	Acetaminophen	Intraparenchymal

Results and applications of MedCAT | NER + L*

Model	Hospital Test Site	# Annotated Examples	F1 μ	F1 SD \pm	F1 IQR
M1: Base - No Training	KCH	3,358	0.638	0.297	0.333
M2: Base + Self-Supervised MIMIC-III	KCH	3,358	0.840	0.109	0.150
M3: Base + Self-Supervised KCH	KCH	3,358	0.889	0.078	0.103
M4: KCH Self-Supervised + KCH Supervised	KCH	3,358	0.947	0.044	0.051
M4: KCH Self-Supervised + KCH Supervised	UCLH	499	0.903	0.103	0.112
M5: KCH Self-Supervised + KCH Supervised + UCLH Self-Supervised	UCLH	499	0.905	0.079	0.034
M6: KCH Self-Supervised + KCH Supervised + UCLH Self-Supervised + UCLH Supervised	UCLH	499	0.926	0.060	0.086
M4: KCH Self-Supervised + KCH Supervised	SLaM	1,425	0.885	0.095	0.088
M7: KCH Self-Supervised + KCH Supervised + SLaM Self-Supervised	SLaM	1,425	0.907	0.047	0.082
M8: KCH Self-Supervised + KCH Supervised + SLaM Self-Supervised + SLaM Supervised	SLaM	1,425	0.945	0.029	0.025

The MedCAT Workflow

1 / 2: Prepare the environment

- Install MedCAT (pip install medcat)
- Prepare datasets
- Choose a vocabulary (existing, or bespoke)
- Choose a Concept DB or Subset
 - UMLS, SNOMED-CT, HPO etc. OR Build a custom one

3: Run unsupervised training

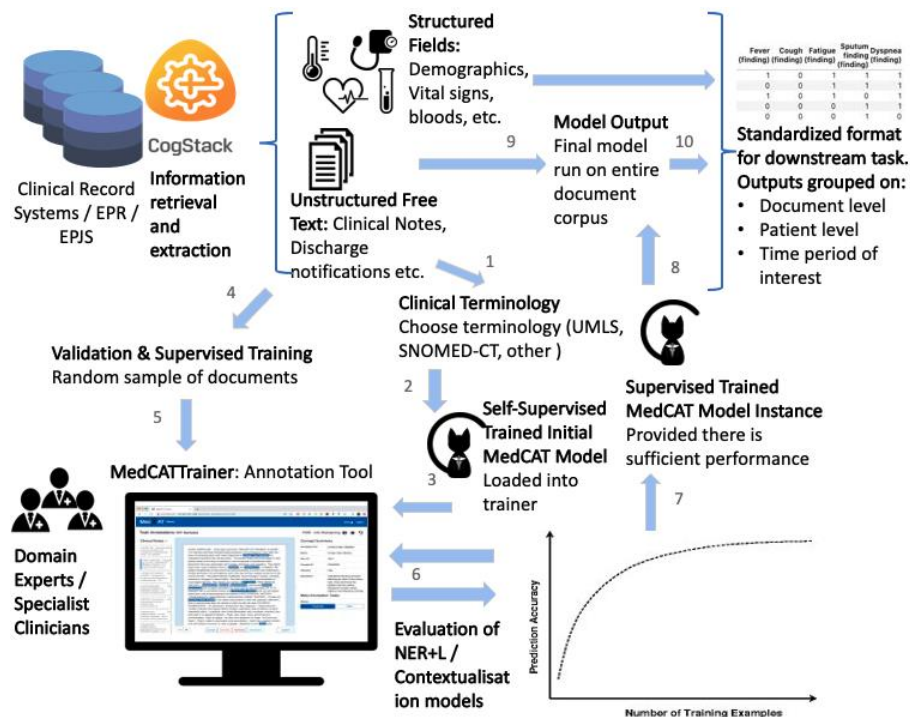
- Run over large biomedical corpora (e.g. a hospital site EPR indexed via CogStack)

4 / 5 / 6: Validate / Fine-tune using MedCATtrainer:

- Validate annotations with clinical teams, build model trust and measure annotation performance.
- Improve annotation performance through supervised training

7 / 8 / 9: Run final models on full dataset:

- Aggregate and structure for downstream use.



Web Application - Annotation Interface

← → ↺ Not Secure | 10.200.104.228:8001/train-annotations/1/412

MedAT Login

Train Annotations: Test

pt_1231_hem | 2 Remaining ?

Clinical Notes

This patient is undergoing 3-dimensionally planned radiation therapy in order to adequately target structures at risk while diminishing the degree of exposure to uninvolved adjacent normal structures. This optimizes the chance of controlling tumor while diminishing the acute and long-term side effects. With conformal 3-dimensional simulation, there is extended physician, therapist, and dosimetrist effort and time expended. The patient is initially taken into a conventional simulator room where appropriate markers are placed and the patient is positioned and immobilized. Preliminary field sizes and arrangements, including gantry angles, collimator angles, and number of fields are conceived. Radiographs are taken and these films are approved by the physician. Appropriate marks are placed on the patient's skin or on the immobilization device.

The patient is transferred to the diagnostic facility and placed on a flat CT scan table. Scans are performed through the targeted area. The scans are evaluated by the radiation oncologist and the tumor volume, target volume, and critical structures are outlined on the CT images. The dosimetrist then evaluates the slices in the treatment-planning computer with appropriately marked structures. This volume is reconstructed in a virtual 3-dimensional space utilizing the beam's-eye view features. Appropriate blocks are designed. Multiplane computerized dosimetry is performed throughout the volume. Field arrangements and blocking are modified as necessary to provide coverage of the target volume while minimizing dose to normal structures.

Once appropriate beam parameters and isodose distributions have been confirmed on the computer scan, the individual slices are then reviewed by the physician. The beam's-eye view, block design, and appropriate volumes are also printed and reviewed by the physician. Once these are approved, physical blocks or multi-leaf collimator equivalents will be devised. If significant changes are made in the field arrangements from the

CC: Seizures.

HX: The patient was initially evaluated at UIHC at 7 years of age. He had been well until 7 months prior to evaluation when he started having spells which were described as dizzy spells lasting from several seconds to one minute in duration. They occurred

Concept Summary

Annotated Text	radiation therapy
Name	Therapeutic radiology procedure
Term ID	T061
Semantic Type	Therapeutic or Preventive Procedure
Concept ID	C1522449
Accuracy	1.00
Description	Treatment of a disease by means of exposure of the target or the whole body to radiation. Radiation therapy is often used as part of curative therapy and occasionally as a component of palliative treatment for cancer. Other uses include total body

Meta Annotation Tasks

Negation

Submit

Negated

Not Negated

Correct Wrong Alternative Concept



CogStack

 KING'S HEALTH PARTNERS

KING'S
College
LONDON

Group lead: Prof. Richard Dobson

Funding: HDR UK, NIHR Maudsley BRC

Publications:

Kraljevic & Searle et al (2020): arxiv.org/abs/2010.01165
Mascio & Kraljevic et al (2020): [dx.doi.org/10.18653/v1/2020.bionlp-1.9](https://doi.org/10.18653/v1/2020.bionlp-1.9)
Bean et al (2020): doi.org/10.1002/ejhf.1924
Carr et al. (2021): doi.org/10.1186/s12916-020-01893-3
Zaker et al(2021): doi.org/10.1016/j.retram.2021.103276
Searle et al. (2020): [dx.doi.org/10.18653/v1/2020.bionlp-1.8](https://doi.org/10.18653/v1/2020.bionlp-1.8)
Searle et al. (2019): [dx.doi.org/10.18653/v1/D19-3024](https://doi.org/10.18653/v1/D19-3024)

GitHub:

- github.com/CogStack
- github.com/CogStack/MedCAT
- github.com/CogStack/MedCATservice
- github.com/CogStack/MedCATtrainer



Team

Zeljko Kraljevic
Daniel Bean
Aurelie Mascio
Lukasz Roguski
Thomas Searle
Amos Folarin
Rebecca Bendayan
James Teo
Richard Dobson



NHS
**National Institute for
Health Research**

King's
College
Hospital **NHS**
NHS Foundation Trust

NHS
**South London
and Maudsley**
NHS Foundation Trust