Creating a Human-Annotated Health Record Dataset with Limited Resources

Krištof Anetta

xanetta@fi.muni.cz

Natural Language Processing Centre

Faculty of Informatics, Masaryk University

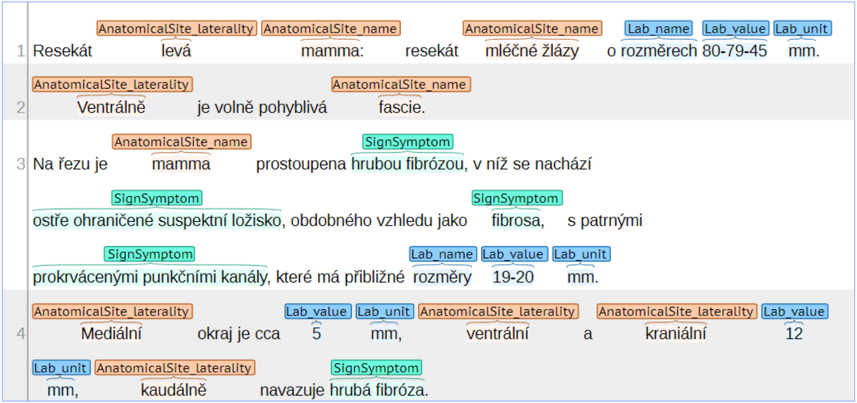
December 9, 2023

Introduction

Introduction

- Getting health records from hospitals is hard
  - Getting annotated ones is next to impossible
- We have a 42-million-word dataset of oncology health records and we are trying to get as much of it annotated
- A balanced set of ca. 50,000 words was chosen to be annotated by humans

Visualization of what we want



1 Resekát levá mamma: resekát mléčné žlázy o rozměrech 80-79-45 mm.

2 Ventrálně je volně pohyblivá fascie.

3 Na řezu je mamma prostoupena hrubou fibrózou, v níž se nachází ostře ohraničené suspektní ložisko, obdobného vzhledu jako fibrosa, s patrnými prokrvácenými punkčními kanály, které má přibližné rozměry 19-20 mm.

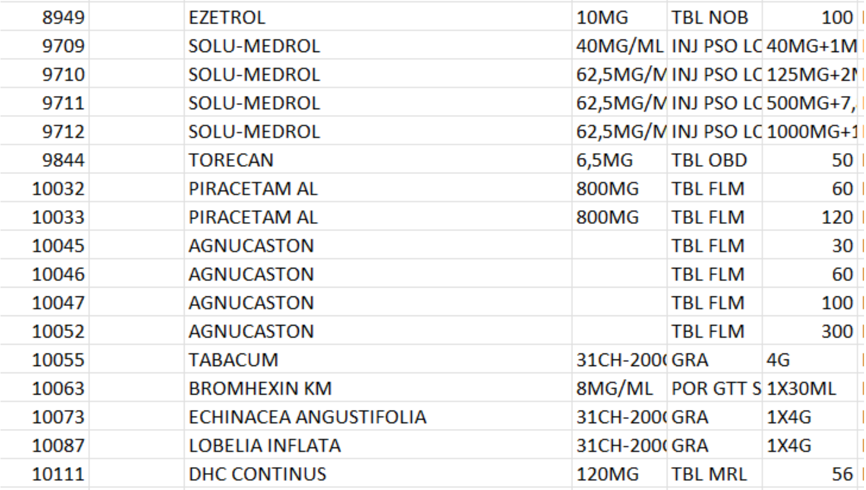
4 Mediální okraj je cca 5 mm, ventrální a kraniální 12 mm, kaudálně navazuje hrubá fibróza.

Before annotation

Preannotation

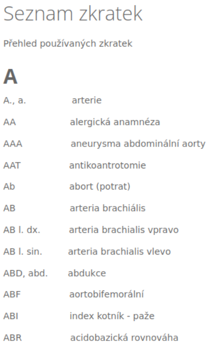
- Easy concepts can be found with rule-based methods
- It is much faster and more accurate when they are already present and annotators only verify them (but they must verify them)
- We preannotated:
    - Medication names
    - Medical abbreviations

Medication: SÚKL databases



8949		EZETROL	10MG	TBL NOB	100
9709		SOLU-MEDROL	40MG/ML	INJ PSO LC	40MG+1M
9710		SOLU-MEDROL	62,5MG/M	INJ PSO LC	125MG+2M
9711		SOLU-MEDROL	62,5MG/M	INJ PSO LC	500MG+7,
9712		SOLU-MEDROL	62,5MG/M	INJ PSO LC	1000MG+1
9844		TORECAN	6,5MG	TBL OBD	50
10032		PIRACETAM AL	800MG	TBL FLM	60
10033		PIRACETAM AL	800MG	TBL FLM	120
10045		AGNUCASTON		TBL FLM	30
10046		AGNUCASTON		TBL FLM	60
10047		AGNUCASTON		TBL FLM	100
10052		AGNUCASTON		TBL FLM	300
10055		TABACUM	31CH-200	GRA	4G
10063		BROMHEXIN KM	8MG/ML	POR GTT S	1X30ML
10073		ECHINACEA ANGUSTIFOLIA	31CH-200	GRA	1X4G
10087		LOBELIA INFLATA	31CH-200	GRA	1X4G
10111		DHC CONTINUS	120MG	TBL MRL	56

Abbreviations: Web resources

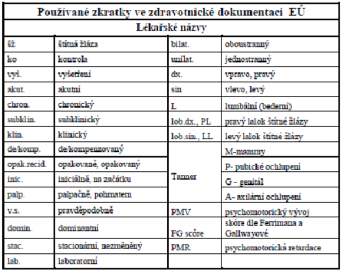


Seznam zkratek

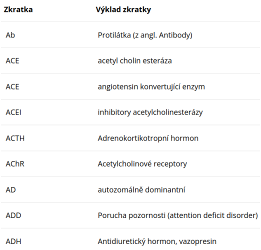
Přehled používaných zkratek

A

A., a.	arterie
AA	alergická anamnéza
AAA	aneurysma abdominální aorty
AAT	antikoantrotomie
Ab	abort (potrat)
AB	arteria brachiális
AB l. dx.	arteria brachialis vpravo
AB l. sin.	arteria brachialis vlevo
ABD, abd.	abdukce
ABF	aortobifemorální
ABI	index kotník - paže
ABR	acidobazická rovnováha



Používané zkratky ve zdravotnické dokumentaci EÚ			
Lékařské názvy			
šž:	štítná žláza	bilat.	oboustranný
ko	kontrola	unilat.	jednostranný
vyš.	vyšetření	dx.	vpravo, pravý
akut.	akutní	sin	vlevo, levý
chron.	chronický	L	lumbální (bederní)
subklin.	subklinický	lob.dx., PL	pravý lalok štítné žlázy
klin.	klinický	lob.sin., LL	levý lalok štítné žlázy
de/komp.	de/kompenzovaný	Tanner	M-mammy
opak.recid.	opakovaně, opakovaný		P- pubické ochlupení
inic.	iniciálně, na začátku		G - genitál
palp.	palpačně, pohmatem		A- axilární ochlupení
v.s.	pravděpodobně	PMV	psychomotorický vývoj
domin.	dominantní	FG scóre	skóre dle Ferrimana a Gallwayové
stac.	stacionární, nezměněný	PMR	psychomotorická retardace
lab.	laboratorní		



Zkratka	Výklad zkratky
Ab	Protilátka (z angl. Antibody)
ACE	acetyl cholin esteráza
ACE	angiotensin konvertující enzym
ACEI	inhibitory acetylcholinesterázy
ACTH	Adrenokortikotropní hormon
AChR	Acetylcholinové receptory
AD	autozomálně dominantní
ADD	Porucha pozornosti (attention deficit disorder)
ADH	Antidiuretický hormon, vazopresin

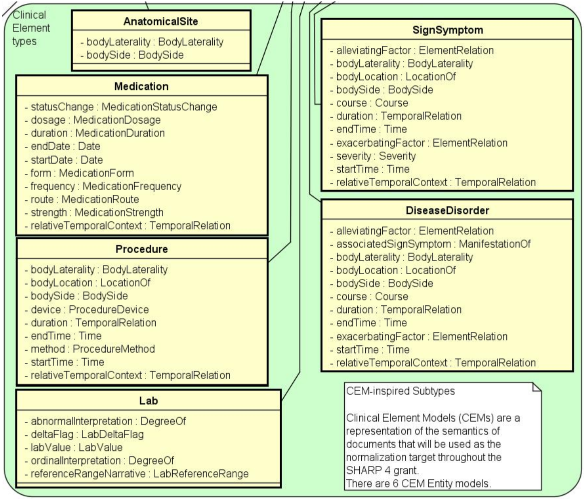
https://nlp.fi.muni.cz/projekty/ehr_analysis/zkratky/

Annotation

Annotators

- 11 students
- Confidence considerations:
  - Since they are not experts, each record is annotated by at least 2 students
  - When they both annotate a string in the same way, confidence is high
  - But even when their views differ, they may both be right - we mark those as 50% confidence

Annotation schema: Apache cTAKES types



Clinical Element types

AnatomicalSite
- bodyLaterality : BodyLaterality
- bodySide : BodySide

Medication
- statusChange : MedicationStatusChange
- dosage : MedicationDosage
- duration : MedicationDuration
- endDate : Date
- startDate : Date
- form : MedicationForm
- frequency : MedicationFrequency
- route : MedicationRoute
- strength : MedicationStrength
- relativeTemporalContext : TemporalRelation

Procedure
- bodyLaterality : BodyLaterality
- bodyLocation : LocationOf
- bodySide : BodySide
- device : ProcedureDevice
- duration : TemporalRelation
- endTime : Time
- method : ProcedureMethod
- startTime : Time
- relativeTemporalContext : TemporalRelation

Lab
- abnormalInterpretation : DegreeOf
- deltaFlag : LabDeltaFlag
- labValue : LabValue
- ordinalInterpretation : DegreeOf
- referenceRangeNarrative : LabReferenceRange

SignSymptom
- alleviatingFactor : ElementRelation
- bodyLaterality : BodyLaterality
- bodyLocation : LocationOf
- bodySide : BodySide
- course : Course
- duration : TemporalRelation
- endTime : Time
- exacerbatingFactor : ElementRelation
- severity : Severity
- startTime : Time
- relativeTemporalContext : TemporalRelation

DiseaseDisorder
- alleviatingFactor : ElementRelation
- associatedSignSymptom : ManifestationOf
- bodyLaterality : BodyLaterality
- bodyLocation : LocationOf
- bodySide : BodySide
- course : Course
- duration : TemporalRelation
- endTime : Time
- exacerbatingFactor : ElementRelation
- startTime : Time
- relativeTemporalContext : TemporalRelation

CEM-inspired Subtypes

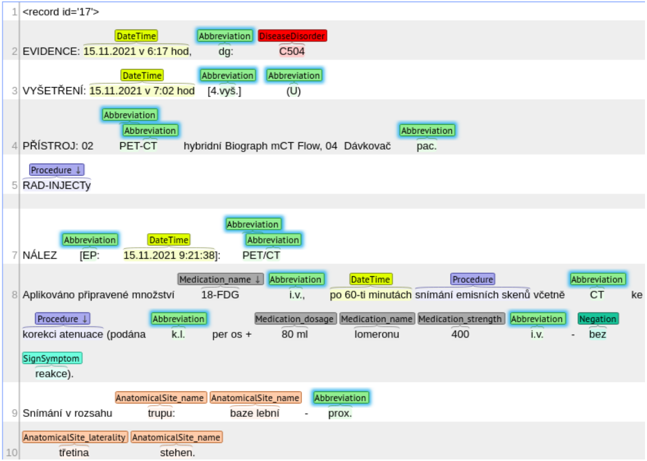
Clinical Element Models (CEMs) are a representation of the semantics of documents that will be used as the normalization target throughout the SHARP 4 grant.
There are 6 CEM Entity models.

The referential semantics schema used in Apache cTAKES [1]

Annotation schema

cTAKES type	Our schema
AnatomicalSite	AnatomicalSite_laterality AnatomicalSite_name
Medication	Medication_dosage Medication_name Medication_strength
Procedure	Procedure
Lab	Lab_name Lab_unit Lab_value
SignSymptom	SignSymptom
DiseaseDisorder	DiseaseDisorder
	DateTime
	Abbreviation
	Negation

Annotation in BRAT



1 <record id='17'>

2 EVIDENCE: 15.11.2021 v 6:17 hod, dg: C504

3 VYŠETŘENÍ: 15.11.2021 v 7:02 hod [4.vyš.] (U)

4 PŘÍSTROJ: 02 PET-CT hybridní Biograph mCT Flow, 04 Dávkovač pac.

5 RAD-INJECTy

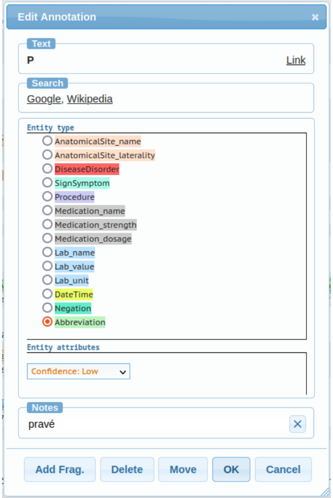
7 NÁLEZ [EP: 15.11.2021 9:21:38]: PET/CT

8 Aplikováno připravené množství 18-FDG i.v., po 60-ti minutách snímání emisních skenů včetně CT ke korekci atenuace (podána k.l. per os + 80 ml Iomeronu 400 i.v. - bez reakce).

9 Snímání v rozsahu trupu: baze lební - prox.

10 třetina stehen.

Annotator options in BRAT



Edit Annotation ×

Text

P Link

Search

Google, Wikipedia

Entity type

○ AnatomicalSite_name
○ AnatomicalSite_laterality
○ DiseaseDisorder
○ SignSymptom
○ Procedure
○ Medication_name
○ Medication_strength
○ Medication_dosage
○ Lab_name
○ Lab_value
○ Lab_unit
○ DateTime
○ Negation
● Abbreviation

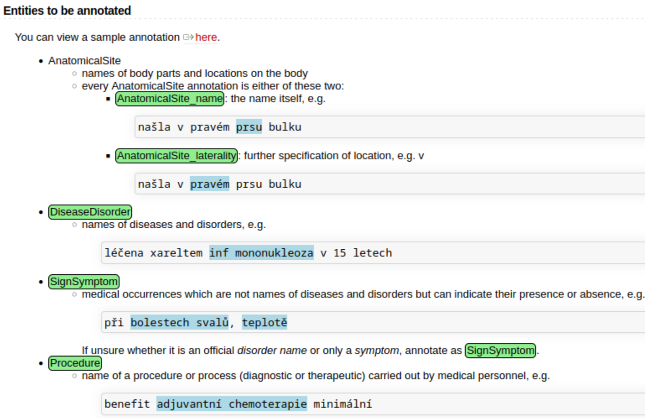
Entity attributes

Confidence: Low

Notes

pravé

Annotator’s manual



Entities to be annotated

You can view a sample annotation ⇨here.

- AnatomicalSite
  - names of body parts and locations on the body
  - every AnatomicalSite annotation is either of these two:
    - AnatomicalSite_name: the name itself, e.g.

      ```
      našla v pravém prsu bulku
      ```

    - AnatomicalSite_laterality: further specification of location, e.g. v

      ```
      našla v pravém prsu bulku
      ```

- DiseaseDisorder
  - names of diseases and disorders, e.g.

    ```
    léčena xareltem inf mononukleoza v 15 letech
    ```

- SignSymptom
  - medical occurrences which are not names of diseases and disorders but can indicate their presence or absence, e.g.

    ```
    při bolestech svalů, teplotě
    ```

    If unsure whether it is an official *disorder name* or only a *symptom*, annotate as SignSymptom.
- Procedure
  - name of a procedure or process (diagnostic or therapeutic) carried out by medical personnel, e.g.

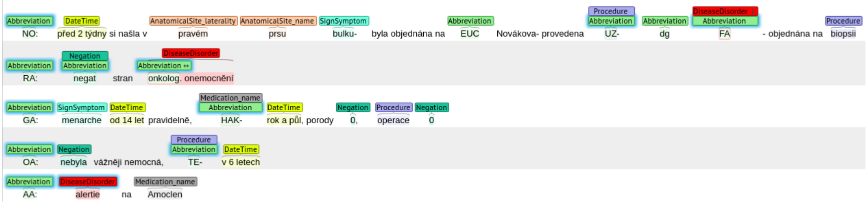
    ```
    benefit adjuvantní chemoterapie minimální
    ```

Annotation statistics

Stage	Annotation count
Initial state of health records	0
Rule-based preannotations	4,266
Preannotations handed to annotators	9,368
New annotations entered by annotators	22,798
Total number of human-verified or human-entered annotations	32,166
Total number of tokens with human-verified or human-entered annotation	45,032

Example of differences

Annotator 1 (record was at the beginning of their dataset)



NO: před 2 týdny si našla v pravém prsu bulku- byla objednána na EUC Novákova- provedena UZ- dg FA - objednána na biopsii

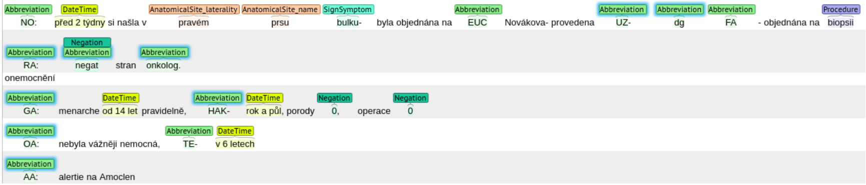
RA: negat stran onkolog. onemocnění

GA: menarche od 14 let pravidelně, HAK- rok a půl, porody 0, operace 0

OA: nebyla vážněji nemocná, TE- v 6 letech

AA: alertie na Amoclen

Annotator 2 (record was in the middle)



NO: před 2 týdny si našla v pravém prsu bulku- byla objednána na EUC Novákova- provedena UZ- dg FA - objednána na biopsii

RA: negat stran onkolog.
onemocnění

GA: menarche od 14 let pravidelně, HAK- rok a půl, porody 0, operace 0

OA: nebyla vážněji nemocná, TE- v 6 letech

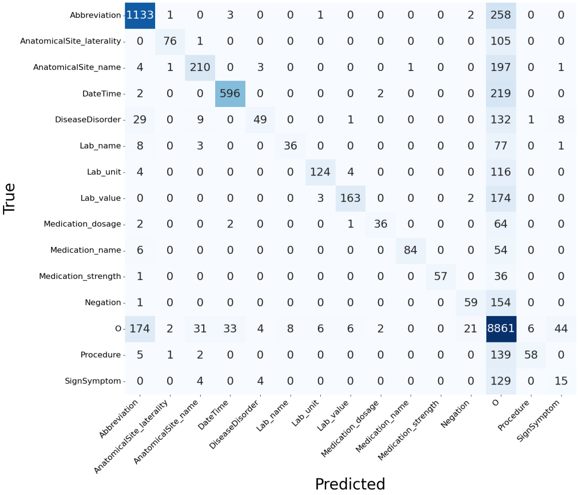
AA: alertie na Amoclen

Preliminary NER training

Preliminary NER training

- Stanford NER
- Cross-evaluation
  - 5 different subsets of data - train on 4, evaluate on 1 as gold standard, calculate average

NER Results



	Abbreviation	AnatomicalSite_laterality	AnatomicalSite_name	DateTime	DiseaseDisorder	Lab_name	Lab_unit	Lab_value	Medication_dosage	Medication_name	Medication_strength	Negation	O	Procedure	SignSymptom
Abbreviation	1133	1	0	3	0	0	1	0	0	0	0	2	258	0	0
AnatomicalSite_laterality	0	76	1	0	0	0	0	0	0	0	0	0	105	0	0
AnatomicalSite_name	4	1	210	0	3	0	0	0	0	1	0	0	197	0	1
DateTime	2	0	0	596	0	0	0	0	2	0	0	0	219	0	0
DiseaseDisorder	29	0	9	0	49	0	0	1	0	0	0	0	132	1	8
Lab_name	8	0	3	0	0	36	0	0	0	0	0	0	77	0	1
Lab_unit	4	0	0	0	0	0	124	4	0	0	0	0	116	0	0
Lab_value	0	0	0	0	0	0	3	163	0	0	0	2	174	0	0
Medication_dosage	2	0	0	2	0	0	0	1	36	0	0	0	64	0	0
Medication_name	6	0	0	0	0	0	0	0	0	84	0	0	54	0	0
Medication_strength	1	0	0	0	0	0	0	0	0	0	57	0	36	0	0
Negation	1	0	0	0	0	0	0	0	0	0	0	59	154	0	0
O	174	2	31	33	4	8	6	6	2	0	0	21	8861	6	44
Procedure	5	1	2	0	0	0	0	0	0	0	0	0	139	58	0
SignSymptom	0	0	4	0	4	0	0	0	0	0	0	0	129	0	15

Preliminary NER training

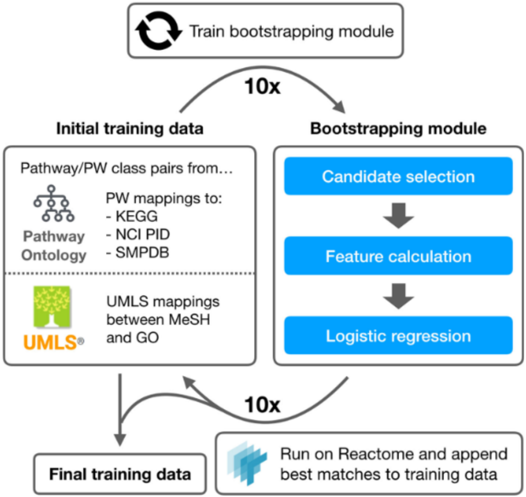
Class	**Precision**	**Recall**	**F1 score**
Abbreviation	0.885	0.707	0.781
AnatomicalSite_laterality	0.871	0.460	0.595
AnatomicalSite_name	0.871	0.426	0.568
DateTime	0.936	0.635	0.753
DiseaseDisorder	0.555	0.235	0.320
Lab_name	0.627	0.297	0.385
Lab_unit	0.778	0.389	0.511
Lab_value	0.832	0.394	0.524
Medication_dosage	0.637	0.286	0.390
Medication_name	0.959	0.554	0.701
Medication_strength	0.822	0.605	0.673
Negation	0.616	0.354	0.416
O (no annotation)	0.803	0.969	0.878
Procedure	0.767	0.319	0.439
SignSymptom	0.440	0.077	0.124
Weighted average	**0.819**	**0.813**	**0.789**

Future directions

Future directions

- 50,000 words is not enough for Transformer training
- Bootstrapping more data:
  - Iterative annotation - train smaller models, annotate larger data, evaluate errors, correct
  - Eventually, the whole 40,000,000 corpus could be annotated - lower quality, but sufficient size for LLMs

Iterative bootstrapping



Train bootstrapping module

10x

Initial training data

Pathway/PW class pairs from…

Pathway Ontology

PW mappings to:
- KEGG
- NCI PID
- SMPDB

UMLS®

UMLS mappings between MeSH and GO

Bootstrapping module

Candidate selection

Feature calculation

Logistic regression

10x

Final training data

Run on Reactome and append best matches to training data

Thank you for your attention!

MUNI
FACULTY
OF INFORMATICS