MUNI
FI

Information extraction from medical records

Tomáš Houfek

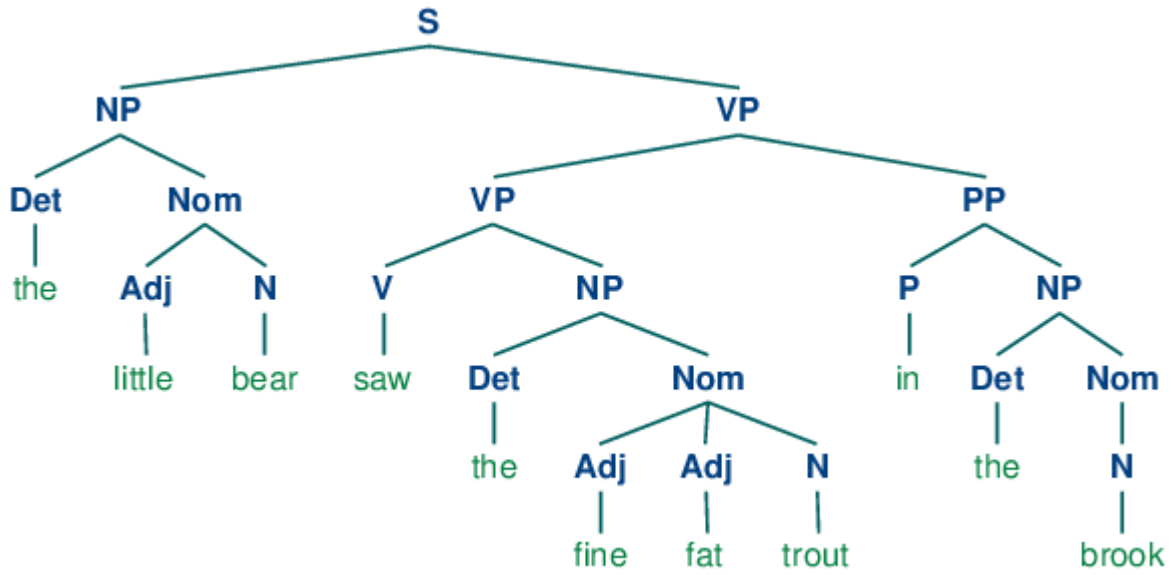
Overview

- General approaches to information extraction
- IE of medical records
- What tools are used for IE in medical records
- IE of Czech medical records
- Problems of information extraction in medical records
- How to approach IE of Czech medical records
- Examples

General approaches

- Pre-processing

- Part of Speech tagging
- Name entity recognition
- Syntactic analysis
- Semantic analysis
- ML
- Regex



MUNI FI

IE in medical records

- Results of 67 studies on IE of medical records
- Majority in US, on US medical records (61%)
- Majority used for detecting cases in medical records (87%)
- Only minority used on hospital EMR (13%)

Tools used for medical records

- Majority used rule-bases NLP algorithms (67%)
- Keyword search (24%)
- Only 9% of ML, Bayesian or hybrid approaches
- A lot of reoccurring IE systems MedLEE (9 studies), cTAKES (5 studies), HITEx (4 studies)

Algorithm accuracy

	No. of Studies	Sensitivity (Recall)	Specificity	PPV (Precision)	Negative predictive value	F measure	AUROC
Algorithm type							
Single algorithm for NLP and case detection	15	96.2	97.4	85.35	96.6	49	–
Rule-based secondary case detection algorithm	20	91.2	95.45	77.5	98.95	97.57	94.4
Probabilistic secondary case detection algorithm (Logistic Regression; Bayesian; machine learning)	21	80	95	86	95.4	77	94

MUNI FI

IE of Czech medical records

- No cTAKES or any other already existing tools
- For every task a specific solution
- Little to none research

MUNI FI

Problem of Czech medical records

- Text of Czech medical records is not typical Czech text
- Lingvistic analysis cannot be used successfully
- Partial solution: Třífázová metoda předběžného zpracování (3PP)
  - Tokenization
  - Normalization
  - Semantic annotation

Extrakce informací z lékařských textů, Ing. Karel Zvára PhD.

MUNI FI

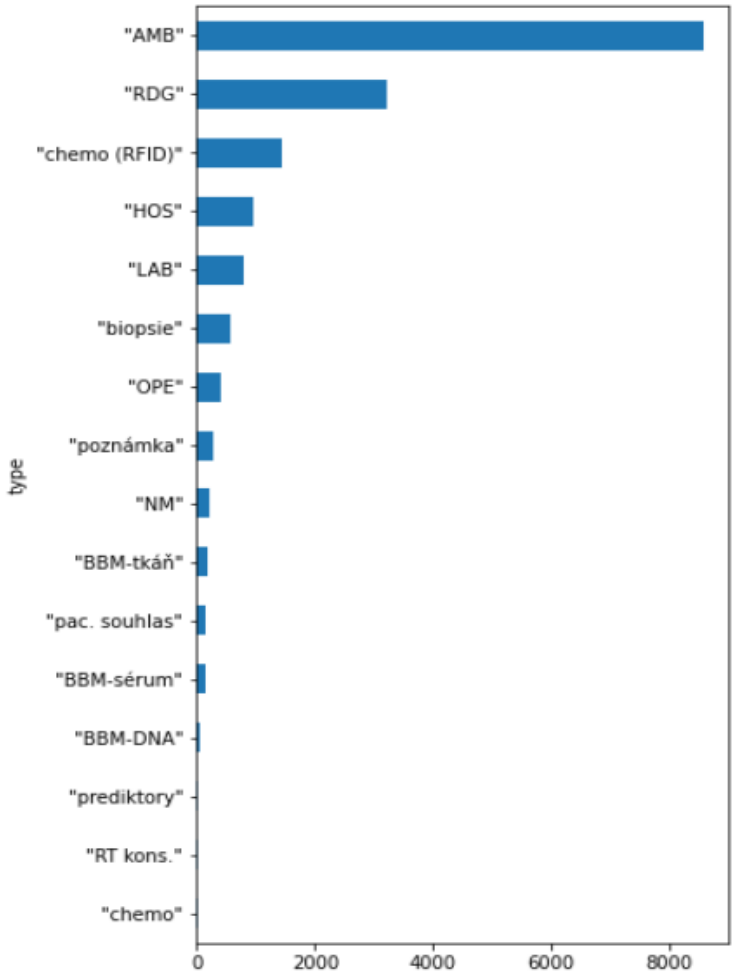
Example of Czech medical record

XX/XX/XX;"AMB";"A-DIG";"MUDr. John Doe";
Pac. s koloskop. prokázaným ca sigmatu v 55 cm
- cirkulární stenosující tu, lymfadenopatie, hepar bez evid. meta,
 na plicích nejasná densita 6 mm - viz dokumentace.
Dop: oper. řešení výrazně stenosujícího tumoru, další dle stagingu pooper.
Příjem OCHIR XX.XX.XX , oper. následující den.

XX/XX/XX;"biopsie";"PAT-B";
Klinická diagnóza (popř. stručný klinický průběh): susp. malignita ve 20cm
Mikroskopický nález: 3x tubulární adenokarcinom infiltrující
v submukóze pod intaktní kolickou sliznicí.
Morfologie odpovídá kolorektálnímu origu.
Závěr: Adenokarcinom v endoskopické excizi z tlustého střeva
KLASIFIKACE: topografie(ICD-O-3): C189, morfologie(ICD-O-3): 8140/39, dg: C189
PROVEDENO: XX.XX.XXXX v 14:52 hod

IE of Czech medical records

- Available data
  - 122 patients
  - 17 216 records
- Types of records – (16)
  - Subtypes of records - (192)



What data I want to extract

- Diagnosis date
- Diagnosis (MNK-10)
- Diagnosis determined by what?
- Morphology
- Laterality

- Treatment in time
  - Operations
  - Chemo

- TNM classification
- pTNM classification
- Lokalization of metastasis
- Clinical stage
- Progression of disease

- State of tumour in time
  - Stadium
  - Size
  - Recurrence

MUNI FI

Example of tagged records

XX/XX/XX;"biopsie";"PAT-B";

Klinická diagnóza (popř. stručný klinický průběh): susp. malignita ve 20cm

(diagnosis word)

Mikroskopický nález: 3x tubulární adenokarcinom infiltrující

(tumor description)

v submukóze pod intaktní kolickou sliznicí.

Morfologie odpovídá kolorektálnímu origu.

Závěr: Adenokarcinom v endoskopické excizi z tlustého střeva

KLASIFIKACE: topografie(ICD-O-3): C189 , morfologie(ICD-O-3): 8140/39 , dg: C189

(topography) (morphology) (diagnosis)

PROVEDENO: XX.XX.XXXX v 14:52 hod

MUNI FI

Example of record tagged

XX/XX/XX;"AMB";"A-DIG";"MUDr. John Doe";

diagnosis word

Pac. s koloskop. prokázaným ca sigmatu v 55 cm

tumor description

- cirkulární stenosující tu, lymfadenopatie, hepar bez evid. meta,
na plicích nejasná densita 6 mm - viz dokumentace.

tumor description

Dop: oper. řešení výrazně stenosujícího tumoru, další dle stagingu pooper.
Příjem OCHIR XX.XX.XX , oper. následující den.

Sources

- Extrakce informací z lékařských textů, Ing. Karel Zvára PhD.
  - http://hdl.handle.net/20.500.11956/94214
- Extracting information from the text of electronic medical records to improve case detection: a systematic review
  - https://pubmed.ncbi.nlm.nih.gov/26911811/
- Data Mining from Free-Text Health Records: State of the Art, New Polish Corpus
  - https://nlp.fi.muni.cz/raslan/2020/paper5.pdf