Clinical language processing: the first steps

Simon Šuster
Madhumita
Stéphan Tulkens
Walter Daelemans

ACCUMULATE

4-year SBO project
5 partners

Exploiting clinical records

Develop technology for analysis of free text of clinical reports in English and Dutch

Goals of Accumulate:

- Terminology extraction
- Analysis of event structure
- Analysis of temporal, spatial and causal information
- Visualization

Exploiting clinical records

Develop technology for analysis of free text of clinical reports in English and Dutch

Work at CLiPS:

- **Terminology extraction**
  - Developing techniques for normalizing the reports
  - Recognizing and disambiguating concepts
- **Analysis of event structure**
  - Capturing predicate-argument structure

Exploiting clinical records

Develop technology for analysis of free text of clinical reports in English and Dutch

Work at CLiPS:

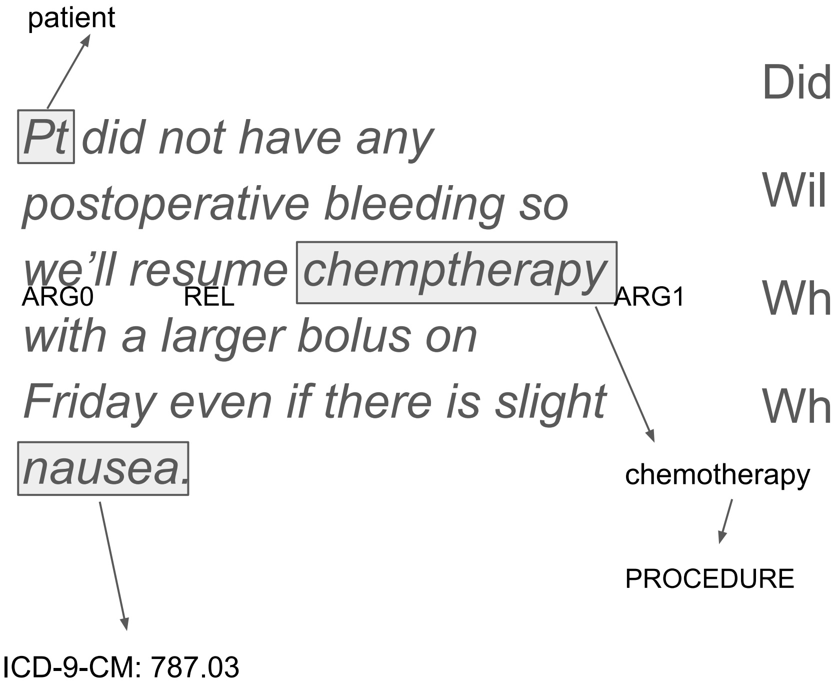
- **Terminology extraction**
  - Developing techniques for normalizing the reports
  - Recognizing and disambiguating concepts
- **Analysis of event structure**
  - Capturing predicate-argument structure

+ Participation at the “Psychiatric symptom severity identification” challenge

We’re often wondering about the impact of our work...

Here, we can make a difference to people’s life!

- Assisting in disease diagnosis (clinical decision support)
- Finding new linkages between symptoms, drugs, diseases, patient attributes…
- Recruiting patients for clinical trials
- Personalized medicine
- Insights into population health



patient

Pt did not have any postoperative bleeding so we'll resume chemptherapy with a larger bolus on Friday even if there is slight nausea.

ARG0

REL

ARG1

chemotherapy

PROCEDURE

ICD-9-CM: 787.03

Did the patient undergo an operation?

Was the patient bleeding?

Did the patient have chemotherapy before?

Will the patient have another round of chemo?

Who could have nausea?

What degree of nausea is acceptable?

Challenges of the clinical domain

- Language characteristics: ill-formedness, abbreviations, acronyms, idiosyncrasies, verb omission
- Poor report structure
- Domain fragmentation
  - paediatrics, cardiology, dermatology, ophthalmology, obstetrics etc.

Why has progress been slow?

- Biomedical- and NLP-community divide
- Little annotated data
- Difficult access to (any) data
    - although plenty

" the problem is not access to annotated data, the problem is access to data [...] we have grad students who are incredibly smart who are working on beer reviews and twitter and emojis because that's where the data is, not because they are not interested in applying [clinical NLP] techniques"

Philip Resnik, NAACL’16

Why is it hard to get access to clinical data?

- Risk of revealing personal information
- Risk of revealing hospital practices (accountability)
- Clinicians lack interest/trust in our work (?)
- Clinicians afraid of losing their unique role (?)

In the meantime...

- DeepMind’s access via UK’s NHS to data of 1.6M patients
- IBM Watson’s access to around 300M of US patient records and patient-generated records (e.g. from wearables)

Do we have oversight of the usage of records and personal data?

What is the science behind?