Clinical data characteristics and processing challenges

19th December 2016

Madhumita Sushil

UZA

Hospital data characteristics

- Limited availability of patient records prior to EMR usage
- Noisy
  - Missing data
    - Unavailable/scanned patient history
    - Patient privacy concerns in sensitive departments like psychiatry
  - EMR format to raw text conversion errors
  - Spelling and typing errors
- Limited availability of structured data for evaluation

Clinical Text Characteristics

- Error prone and noisy
- Abbreviated and coded language use
  - Abbreviations and acronyms
  - Symptom presence: Stress+, Stress- (Hyphen or negation?)
  - Symptom intensity indicators
    - Stress++
    - >stress
  - Approximation: +/- 4 weken, +- 1 week
  - Comparison/Interval indicator: >10u

Clinical Text Characteristics

- Dutch compounding: possibly very long
  - Example: *breedspectrumantibioticabehandeling*
- Language switching
  - Certain terms in e.g. English, French
  - Complete documents in different language

Points for Discussion

Developing transferable technology

- Question: Adapt data to tools or tools to data?
  - Data adaptation: Text cleanup – how much?
    - Losing data-specific properties
    - Limiting transferability to real-world applications
    - Cleaned data quality?
  - Tool adaptation
    - Retrain modules on clinical data – data annotation requirements

Developing transferable technology

- Assuring language independence
  - Using common resources/ontologies
  - Making languages compatible
    - Decompound Dutch terms?
    - Different word orders?

Data annotation within Accumulate

- What to annotate?
  - Linguistic: Tokens, POS tags, syntactic structure
  - Medical concepts and their types
  - Negation, modality and their scope
  - Temporal entities and relations
  - Spatial entities and relations
  - Relations between medical concepts
  - Abbreviations and acronyms with full forms
  - Spelling errors

Data annotation within Accumulate

- Which documents to annotate?
- How much to annotate?
- Which guidelines to follow?
  - English vs. Dutch guidelines

Valorization

- Application requirements
  - High precision – minimum misleading information
  - Goal: support systems for medical professionals
    - Outside clinical workflow – minimize quality control requirements
    - Minimum hassle for them – ensure usability

Valorization

- Some valorization potentials
  - Efficient clinical data visualization
  - Patient profiling and recruitment for clinical trials
  - Improving clinical hypothesis generation using mined data – exploratory analysis
  - Clinical NLP tools