Clinical text retrieval - some methods and some applications

Hercules Dalianis
Clinical Text Mining group
Department of Computer and Systems Sciences (DSV)
hercules@dsv.su.se

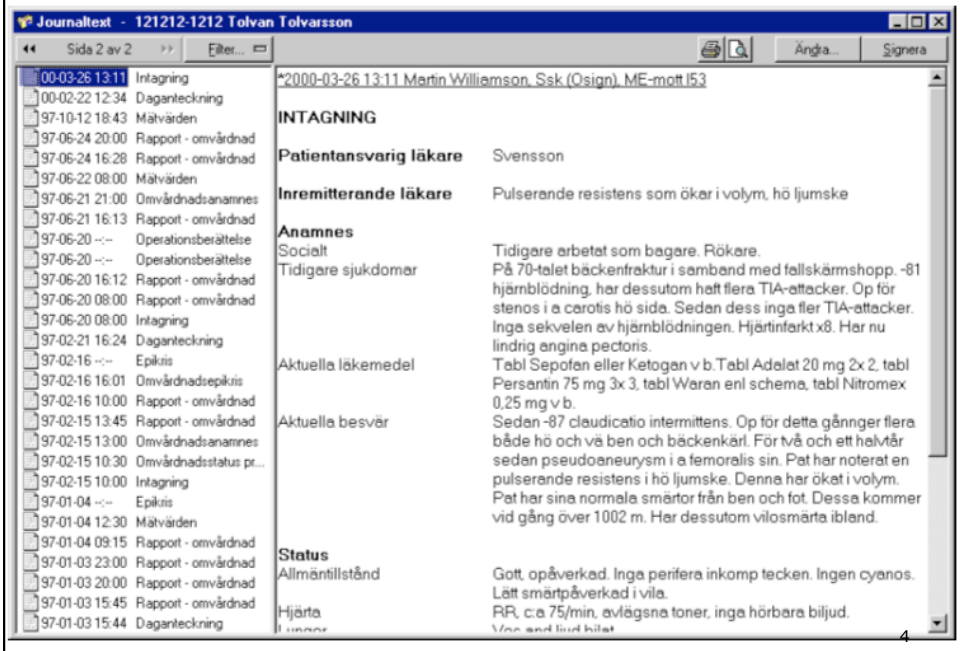
Overview

- Background about clinical data
- Symptoms, diagnoses, drugs, bodyparts
- Diagnosis Finder (Supervised)
- Text mining ICD diagnosis codes (Unsupervised)
- Hospital aquired infections

Stockholm University

Why clinical text mining

- 4-10 million pages of patient records are produced each year in Sweden (pop. 10 million)
- The records contain both structured data and unstructured data - free text



Stockholm EPR Corpus

- More one million in-patients
- Year 2006-2010
- From Karolinska University Hospital
- De-identified but still sensitive
- 500 clinics/units

Content in patient records

- Serial number, gender, age
- Admission, discharge date and time
- Blood-, laboratory values, ICD-10 diagnosis codes
- Drugs ATC-codes
- Free text in Swedish
  - Physician's notes, reasoning, nurses narratives, etc
- Ethical permissions!!

Example record (Anonymized manually)

123 H - IVA 322916614D 2007-08-21 9:12
1944 Kvinna Anamnesis

Kvinna med hjrtsvikt, förmaksflimmer, angina pectoris. Ensamstående änka. Tidigare CVL med sequelae högersidig hemipares och afasi. Tidigare vårdad för krampanfall misstänkt apoplektisk. Inkommer nu efter att ha blivit hittad på en stol och sannolikt suttit så över natten. Inkommer nu för utredning. Sonen Johan är med.

23 H - IVA 322916614D 2008-08-21 10:54
1944 Kvinna Bedömning

Grav hjärtsvikt efter hjärtinfarkt x 2 inklusive eoisod med asystoli och HLR. EF 20-25%. Neurologisk påverkan med hösidig svaghet. Blodprov. Odlingar tas i blod och urin. Remiss skickas pulm-rtg enl dr Svenssons anteckning. Atelektaser. Pneumoni, I110. Hjärtinsufficiens, ospecificerad, I509

(English translation)

123 H - IVA 322916614D 2008-08-21 9:12

1944 Woman Anamnesis

Woman with hert failures, atrial fibrillation, and angina pectoris. Single widow. Former CVL with sequele, rght hemiparesis and aphasia. Prior hospital care for seizures, suspected to be apoepeleptic. Arrive to hospital after being found in a chair and probably been sitting there over night. Arrive for further investigation and care. Accompanied by her son Johan.

123 H - IVA 322916614D 2008-08-21 10:54

1944 Woman Assessment/Plan

Severe heart failure after heart infarction x 2. including episode with heart arrest and acute heart arrest treatment. Ejection fracture (EF) 20-25%. Neurological symptoms with right sided hemiparesis. Blood samples. Culture for blood and urine. Referral for pulmonary x-ray according to dr Svensson's notes. Atelectases. Pneumonia, I110. Heart failure, unspecified, I509.

Medicinskt journalspråk

Septisk pat, oklart fokus,
rundodlas före Zinacef

=>

Patienten har sepsis med oklart ursprung, bakterieodling tas från samtliga möjliga infektionsfokus, inklusive blododling, innan behandling med Zinacef inleds.

Medical language

Septicemic pat, unclear origin,
roundcultured before Zinacef.

=>

The patient has septicemia of unclear origin, bacterial culture samples taken from all possible foci for infection, including blood culture samples, before commencing treatment with Zinacef.

Clinical text genre

- Incomplete sentences
- No use of subject, *har ont, "have pain"*
- *Passive verb, krampar, "cramps"*
- Non standards abbreviations, *pat, p5*
  - *Patient, patological*
  - Pertrokantär FEMurfraktur p5
- Misspellings, *Parkisons,*

Detection clinical entities

- Program modules for detection of
  - Symptom and diagnosis
  - Negation
  - Uncertainty
  - Period of time

    76-årig kvinna med hypertoni och angina pectoris. Trolig hjärtinfarkt 2 år sedan. Inkommer med centrala bröstsmärtor utan utstrålning.

  - 76-year old woman with hypertension and angina pectoris. Possible heart attack 2 years ago. Admitted to hospital with central chest pain without radiation.

Detection clinical entities

Stockholm University

- Program modules for detection of
  - Symptom and diagnosis
  - Negation
  - Uncertainty
  - Period of time

76-årig kvinna med hypertoni och angina pectoris. Trolig hjärtinfarkt 2 år sedan. Inkommer med centrala bröstsmärtor utan utstrålning.

- 76-year old woman with hypertension and angina pectoris. Possible heart attack 2 years ago. Admitted to hospital with central chest pain without radiation.

Detection clinical entities

Stockholm University

- Program modules for detection of
  - Symptom and diagnosis
  - Negation
  - Uncertainty
  - Period of time

76-årig kvinna med hypertoni och angina pectoris. Trolig hjärtinfarkt 2 år sedan. Inkommer med centrala bröstsmärtor utan utstrålning.

- 76-year old woman with hypertension and angina pectoris. Possible heart attack 2 years ago. Admitted to hospital with central chest pain without radiation.

Detection clinical entities

Stockholm University

- Program modules for detection of
  - Symptom and diagnosis
  - Negation
  - Uncertainty
  - Period of time

76-årig kvinna med hypertoni och angina pectoris. Trolig hjärtinfarkt 2 år sedan. Inkommer med centrala bröstsmärtor utan utstrålning.

- 76-year old woman with hypertension and angina pectoris. Possible heart attack 2 years ago. Admitted to hospital with central chest pain without radiation.

Detection clinical entities

Stockholm University

- Program modules for detection of
  - Symptom and diagnosis
  - Negation
  - Uncertainty
  - Period of time

76-årig kvinna med hypertoni och angina pectoris. Trolig hjärtinfarkt 2 år sedan. Inkommer med centrala bröstsmärtor utan utstrålning.

- 76-year old woman with hypertension and angina pectoris. Possible heart attack 2 years ago. Admitted to hospital with central chest pain without radiation.

Detection clinical entities

Stockholm University

- Program modules for detection of
  - Symptom and diagnosis
  - Negation
  - Uncertainty
  - Period of time

  76-årig kvinna med hypertoni och angina pectoris. Trolig hjärtinfarkt 2 år sedan. Inkommer med centrala bröstsmärtor utan utstrålning.

  - 76-year old woman with hypertension and angina pectoris. Possible heart attack 2 years ago. Admitted to hospital with central chest pain without radiation.

Stockholm University

Supervised "Diagnosis finder"

- One annotator and one extra for IAA
- **Annotations classes**
  - Finding (Symptom)
  - Disorders (Diagnosis)
  - Drug
  - Body structure

Annotations classes and (instances)

– Findings (Symptom)	2 540
– Disorders (Diagnosis)	1 317
– Drug	959
– Body structure	497

Conditional random fields

- CRF++
- Pre-processing
  - Lemmatisation
  - Terminology matching (SNOMED CT)

Machine learning with CRF ++

Evaluation on held out data

	Precision	Recall	F-score
Disorder	0.80 (± 0.03)	0.82 (± 0.03)	0.81
Finding	0.72 (± 0.03)	0.65 (± 0.03)	0.69
Drug	0.95 (± 0.02)	0.83 (± 0.03)	0.88
Body structure	0.88 (± 0.04)	0.82 (± 0.05)	0.85
Disorder+Finding	0.80 (± 0.02)	0.76 (± 0.02)	0.78

ICD-code errors

- ICD-coding completely manually, low quality and costly
  - 16 000 ICD codes, 22 Sub chapters
  - Time-consuming to assign codes
- 17% missing ICD-10 codes (in our data)
- 20% wrong codes (Socialstyrelsen, The National Board of Health and Welfare)
- Expensive: 25 billion USD/year in U.S. (Lang, 2007)
- Text mining the right codes?

Stockholm University

Unsupervised Automatic ICD-10 code assignment using Random Indexing

- Index all available patient records using random indexing incl. the previously assigned ICD-10 codes to create a word space model.
- Exploits the relation between words and diagnosis codes in a set of record texts
- Related/Associated words are grouped with the ICD-10 codes
- Enter a diagnosis and one gets an ICD-10-code

Stockholm University

Example, ICD-10 code assignments

hosta (cough)

J18.9 - Pneumoni, ospecificerad (Pneumonia, unspecified)
J15.9 - Bakteriell pneumoni, ospecificerad (Bacterial pneumonia, unspecified)
H66.9 - Mellanöreinflammation, ej specificerad som varig / icke varig (Otitis media, unspecified)
J20.9 - Akut bronkit, ospecificerad (Acute bronchitis, unspecified)
B34.9 - Virusinfektion, ospecificerad (Viral infection, unspecified)
G96.9 - Sjukdom i centrala nervsystemet, ospecificerad (Disorder of central nervous system, unspecified)

=> 82 percent correct assigned codes in Rheumatology

Why: ICD-10 assignment

- Users
  - Physician
    - To assign ICD-10 codes
  - Hospital management
    - To validate ICD-10 codes

Hospital acquired infections: Statistics

- International studies have found that up to 10 per cent of patients at any given time has hospital acquired infections, (Humphreys and Smyths, 2006)
- 10 per cent or more of the in-patients obtain a HAI in Europe
- Three million injured patients and 50 000 deaths yearly only in Europe.

Definition of Hospital Acquired Infection

[a]n infection occurring in a patient in a hospital or other health care facility in whom the infection was not present or incubating at the time of admission. This includes infections acquired in the hospital but appearing after discharge, and also occupational infections among staff of the facility

Hospital Acquired Infection

- It should occur after 48 hours at the ward/ hospital
- It can also occur earlier if the patient has been moved between wards, or at short stay at home less than 24 hours

Types of Hospital Acquired Infections

- pneumonia
- urinary tract infection
- sepsis
- wound infections
- catheter infection
- winter vomiting disease
- etc

Monitoring HAIs

- Compulsory manual reporting by personnel
  - However seldom carried out
- Point Prevalence Measures (PPM)
  - Manual and carried out twice a year (during a day)

HAI Definition is vague

- Definition is vague, who has obtained a HAI?
- From where have they obtained the infection?
- Patients are very ill, multiple sick and in bad shape and become therefore easily infected

Manual monitoring

- Difficult
- Tiresome
- Low IAA by physicians
- Only on a small sample 1-2 percent of all in-patients

Automatic HAI monitoring

- To ease burden of clinicians
- To assist hospital management
- To get better reporting on a larger population

A Hospital Acquired Infection Case

123 H - IVA 322916614D 2007-08-21 9:12
1944 Woman Anamnesis

Pneumonia, I110. Heart failure, unspecified, I509.

Got a urine catheter two days ago. Has now fever. Done a lab test on the urine and gave antibiotics, Penomax.

123 H - IVA 322916614D 2007-08-22 16:12
1944 Woman

No fever. The lab test on urine shows that she had bacteria in the urine.

Information written in the patient record but also in the structured fields for temperature, drugs and lab results.

Temporal and negation

Pat. op. för två dagar sedan

The pat. was op. two days ago

Hon har inte feber, men mycket röd runt op. ställe

She does not have fever, but very red around op. place

Temporal and negation

Pat. op. för två dagar sedan

The pat. was op. two days ago

Hon har inte feber, men mycket röd runt op. ställe

She does not have fever, but very red around op. place

Temporal and negation

Pat. op. för två dagar sedan

The pat. was op. two days ago

Hon har inte feber, men mycket röd runt op. ställe

She does not have fever, but very red around op. place

Temporal and negation

Pat. op. för två dagar sedan

The pat. was op. two days ago

Hon har inte feber, men mycket röd runt op. ställe

She does not have fever, but very red around op. place

Two approaches in Detect-HAI

- Text processing
  - Rule based approach
  - Machine learning based approach

Machine learning based approach

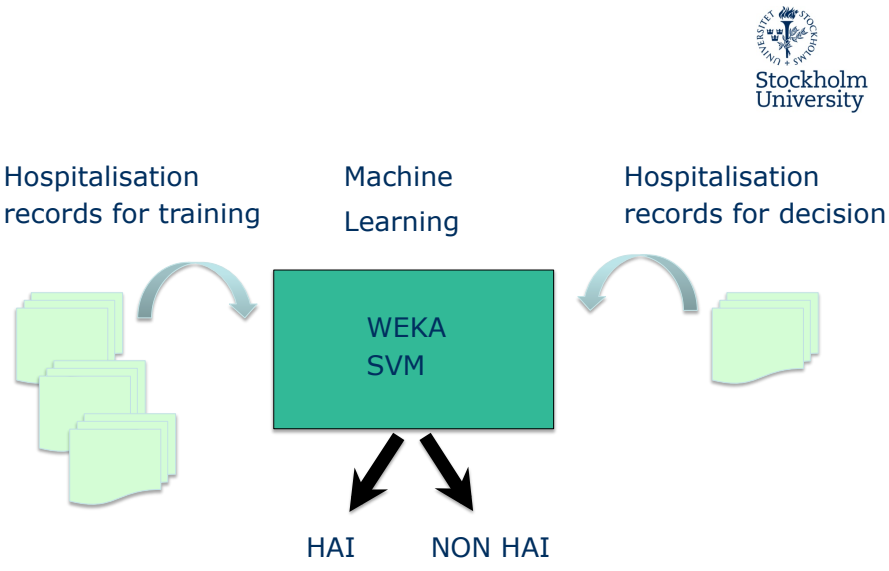
- 215 hospitalisation records (vårdtillfällen)
  - 128 with HAI 1 300 000 tokens
  - 85 without HAI 300 000 tokens
- WEKA Machine learning toolkit using the SVM, Support Vector Machine Algorithm

Stockholm University

IST infection specific terms
1,045 terminology entries

- CT (Computed tomography), kateter (catheter), dränage (drainage), sårinfektion (wound infection), intubering (intubation), operation (surgery), röd (red), urinstämma (urinary retention), ultraljud (ultrasound), feber (fever), . . .

Stockholm University



Stockholm University

Results detecting HAI

- With SVM Support Vector Machine algorithm 89.8% recall and 66.9% precision (See Ehrentraut et al 2012).

Stockholm University

	Naïve Bayes			SVM			C4.5		
	Precision	Recall	F-score	Precision	Recall	F-score	Precision	Recall	F-score
No preprocessing	76.2	60.2	67.2	76.9	**80.5**	78.6	71.4	78.1	74.6
Lemmatized	55.4	**72.9**	62.9	60.0	60.0	60.0	62.2	60.0	61.1
No stop words	78.1	58.6	67.0	74.8	**78.9**	76.8	74.0	75.8	74.9
IST	76.9	62.5	69.0	70.5	**67.2**	68.8	68.0	64.8	66.4
TF-IDF 50	66.4	78.9	72.1	66.9	**89.8**	76.7	68.3	85.9	76.1
LS-TFIDF	67.6	76.6	71.8	63.4	81.3	71.2	66.0	**82.0**	73.2

Rule based approach

- Event
- Device (IN/OUT) VRI-Symptom
- Action
- Antimicrobial Treatment
- Microbiological agent
  - bacteria
  - virus
  - fungi
- VRI-Diagnosis (infection)
- Risk

Rule based approach

- Infections specific types - urinary tract infection
  - Antibiotics
  - bacteria in urine
  - fever
  - and some text mining for *catheter*
- *Results*
  - 80% recall and 87% precision
  - (90% recall and 67% precision for ML on all)

Template for extracted data from tables

Mall för utdata extraherade från tabeller

↓ Mall start ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓

@@@@@|patientnr|kon|fodelsear|handelsedatum|veckodag|@@@@@

<<<<<Journalanteckning>>>>>

#####|journalanteckning_id|vardenhet|yrke|mall|#####

%%%%%%|sokord_term|vardeterm|%%%%%%

ICD-10 kod|kod text

or

anteckning

%%%%%%|sokord_term|(1)vardeterm(2)vardeterm(3)vardeterm...|%%%%%%

ICD-10 kod|kod text

or

anteckning

....

<<<<<Läkemedelsmodul>>>>>

#####|lakemedel_id|#####

ATC-kod|kod text

....

<<<<<Mikrobiologiska Svar>>>>>

#####|svar_uid|undersokning|#####

analysnamn

#####|svar_uid|undersokning|#####

(1)analysnamn

(2)analysnamn

....

<<<<<Kroppstemperatur>>>>>

kroppstemperatur

....

↑ Mall slut ↑

A hospitalisation record

hospitalisation records

@@@@@|011|M|1947|2012-04-29|tisdag|@@@@@

<<<<<Journalanteckning>>>>>

#####|25608293|H - Akutmott (Inf)|Läkare|Intagningsanteckning|#####

%%%%%%|Tid/nuv.sjukdomar|-----|%%%%%%

Välkänd pat på lungklin. Har emfysem och bronkiektasier sedan unga år. Senaste halvåret haft växt av pseudomonas i sputumodl vid upprepade tillfällen och pat har fått upprepade kurer med bredspektrumantibiotika, Tazocin + Meronem. Senaste kuren avslutad den 15/4 och man satte i stället in honom på Azitromax.

%%%%%%|Aktuella läkemedel|-----|%%%%%%

t Calcichew D3 1 x 2

t Betapred 05mg 5 x 1 i nedtrappande dos,

#####|14941941|Blododling, aerob och anaerob|#####

Ingen växt

<<<<<Kroppstemperatur>>>>>

38

38

38,5

@@@@@|011|M|1947|2012-04-30|onsdag|@@@@@

.....

Conclusions of Detect-HAI

- Lower percentage than physician
- But consequent analysis, (physians low IAA)
- 100 per cent analysis on all records 24/7

Summary

- Large amount of unstrured clinical text
- Detection of clinical entities, symptoms, disorders, body parts and drugs
- Assignment and validation of ICD-10 codes
- Detection of Hospital Acquired Infections

Conclusions

- Lots of unstructured text with valuable information
- Large growing repositories saved since long time
- Heavy burden on health care
  - Need to create tools for the clinicians
    - Extraction of information
    - Spell checkers
- Reporting ICD-10 codes
- Reporting and predicting Hospital Acquired Infections
    -

Master thesis work

- http://dsv.su.se/polopoly_fs/1.149704.1383558319!/menu/standard/file/Master%20Thesis-Proposal-Clinical-text-mining-group-2013.pdf
- Clinical Text Mining group
- http://dsv.su.se/en/research/research-areas/health/clintextgroup

Discussion / Questions

Research projects

- Detect-HAI - Detection of Hospital Acquired Infections through language technology analysis of electronic patient records.
  (In Swedish: Detektion av vårdrelaterade infektioner genom språkteknologisk analys av elektroniska patientjournaler)
- High-Performance Data Mining for Drug Effect Detection
  (in Swedish: DataAnalys för DEtektion av Läkemedelseffekter, DADEL)

- NIASC-Nordic Center of Excellence-The Nordic Information for Action eScience Center.
- Automated translation of radiology reports into general Swedish – part of the democratization process in health care
  (in Swedish: Automatiserad översättning av röntgensvar till allmänsvenska - ett led i demokratiseringen av sjukvården)

Detect HAI

- 3 000 deaths and 100 000 patient injuries yearly in Sweden due to adverse events.
- How can we detect and prevent adverse events, specifically Hospital Acquired Infections?
- Use text mining to automatically find the triggers, both in structured and unstructured patient record texts.
- Project time 2012-2014
- Budget 2.2 MSEK

DADEL-High-Performance Data Mining for Drug Effect Detection

- AstraZeneca, WHO-Uppsala Monitoring Centre, Karolinska University hospital, University of Borås
- Drug effect and (new) side effect detection specific in Heart Diseases
- Analyzing patient records, drug registries, case safety reports and chemical compound data in the form of both structured and unstructured (free text) data.
- Connect to registries and bio banks
- Budget: 19 MSEK and five years, 2012-2016

NIASC-Nordic Center of Excellence-The Nordic Information for Action eScience Center

- 13 partners in the Nordic countries, Estonia, and Poland
- E-Science, work with health registries, bio banks and patient records
  Domain cancer and cancer screening.
- Duration 2014-2018
- 44 MSEK Nordforsk

Automated translation of radiology reports into general Swedish – part of the democratization process in health care

- Karolinska University Hospital, Uppsala University, Centrum för lättläst.
- Simplify radiology reports for laymen patients
- Project time 2013-2014
- 1.5 MSEK Vårdalsstiftelsen

References

- Humphreys, H., & Smyth, E. T. M. (2006). Prevalence surveys of healthcare-associated infections: what do they tell us, if anything?. *Clinical Microbiology and Infection*, *12*(1), 2-4.
- Socialstyrelsen. (2010). The National Board of Health and Welfare, Kodningskvalitet i patientregistret, Slutenvård 2008, http://www.socialstyrelsen.se/Lists/Artikelkatalog/Attachments/18082/2010-6-27.pdf
- Skeppstedt, M., Kvist, M., Nilsson, G. H., & Dalianis, H. (2014). Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of biomedical informatics*.