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

Stockholm University

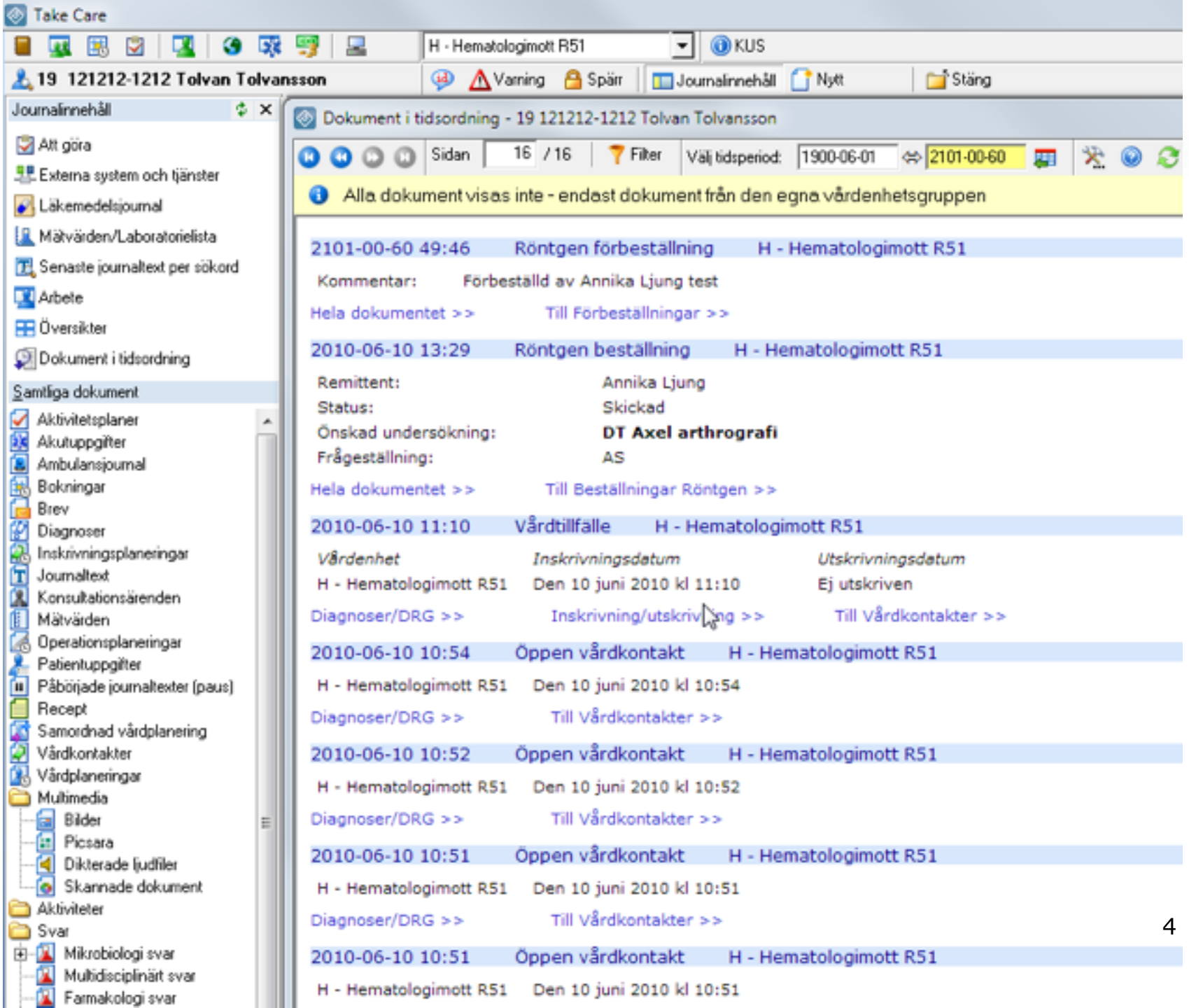
Overview

- Background about clinical data
- Symptoms, diagnoses, drugs, bodyparts
- Diagnosis Finder (Supervised)
- Text mining ICD diagnosis codes (Unsupervised)
- Healthcare Associated Infections (HAI)

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


Take Care
H - Hematologimott R51
KUS
19 121212-1212 Tolvan Tolvansson
Varning
Spärr
Journalinnehåll
Nytt
Stäng
Journalinnehåll
Att göra
Externa system och tjänster
Läkemedelsjournal
Mätvärden/Laboratorielista
Senaste journaltext per sökord
Arbete
Översikter
Dokument i tidsordning
Samtliga dokument
Aktivitetsplaner
Akutuppgifter
Ambulansjournal
Bokningar
Brev
Diagnoser
Inskrivningsplaneringar
Journaltext
Konsultationsärenden
Mätvärden
Operationsplaneringar
Patientuppgifter
Påbörjade journaltexter (paus)
Recept
Samordnad vårdplanering
Vårdkontakter
Vårdplaneringar
Multimedia
Bilder
Picsara
Dikterade ljudfiler
Skannade dokument
Aktiviteter
Svar
Mikrobiologi svar
Multidisciplinärt svar
Farmakologi svar
Dokument i tidsordning - 19 121212-1212 Tolvan Tolvansson
Sidan 16 / 16
Filter
Välj tidsperiod: 1900-06-01 2101-00-60
Alla dokument visas inte - endast dokument från den egna vårdenhetsgruppen
2101-00-60 49:46 Röntgen förbeställning H - Hematologimott R51
Kommentar: Förbeställd av Annika Ljung test
Hela dokumentet >> Till Förbeställningar >>
2010-06-10 13:29 Röntgen beställning H - Hematologimott R51
Remittent: Annika Ljung
Status: Skickad
Önskad undersökning: DT Axel arthrografi
Frågeställning: AS
Hela dokumentet >> Till Beställningar Röntgen >>
2010-06-10 11:10 Vårdtillfälle H - Hematologimott R51
Vårdenhet Inskrivningsdatum Utskrivningsdatum
H - Hematologimott R51 Den 10 juni 2010 kl 11:10 Ej utskriven
Diagnoser/DRG >> Inskrivning/utskrivning >> Till Vårdkontakter >>
2010-06-10 10:54 Öppen vårdkontakt H - Hematologimott R51
H - Hematologimott R51 Den 10 juni 2010 kl 10:54
Diagnoser/DRG >> Till Vårdkontakter >>
2010-06-10 10:52 Öppen vårdkontakt H - Hematologimott R51
H - Hematologimott R51 Den 10 juni 2010 kl 10:52
Diagnoser/DRG >> Till Vårdkontakter >>
2010-06-10 10:51 Öppen vårdkontakt H - Hematologimott R51
H - Hematologimott R51 Den 10 juni 2010 kl 10:51
Diagnoser/DRG >> Till Vårdkontakter >>
2010-06-10 10:51 Öppen vårdkontakt H - Hematologimott R51
H - Hematologimott R51 Den 10 juni 2010 kl 10:51

HEALTH BANK - Swedish HEALTH Record Research Bank (Stockholm EPR Corpus)

- More than two million in-patients
- Year 2006-2014
- From Karolinska University Hospital
- De-identified but still sensitive
- 500 clinics/units
- 23 000 users (readers and writers),
- 6-7 different professions

Content in patient records

- Serial number, gender, age
- Admission, discharge date and time stamps
- 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, pathological*
  - Pertrokantär FEMurfraktur p5
- Misspellings, *Parkisons*
- *Negations ingen feber, "no fever"*
- *Uncertain expressions*
  - *possible Parkisons, propably pneumonia*

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

- 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

- 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

- 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

- 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

- 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.

Supervised "Diagnosis finder"

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

Stockholm University

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-Diagnosis code errors

- ICD-10-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?

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

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

Healthcare associated infections (HAIs): Statistics

- International studies have found that up to 10 per cent of patients at any given time has Health care associated 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 Health care Associated Infection (HAI)

[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

Health care Associated 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 Health care Associated 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 Healthcare Associated 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

Stockholm University

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 and RF, Random Forest algorithms

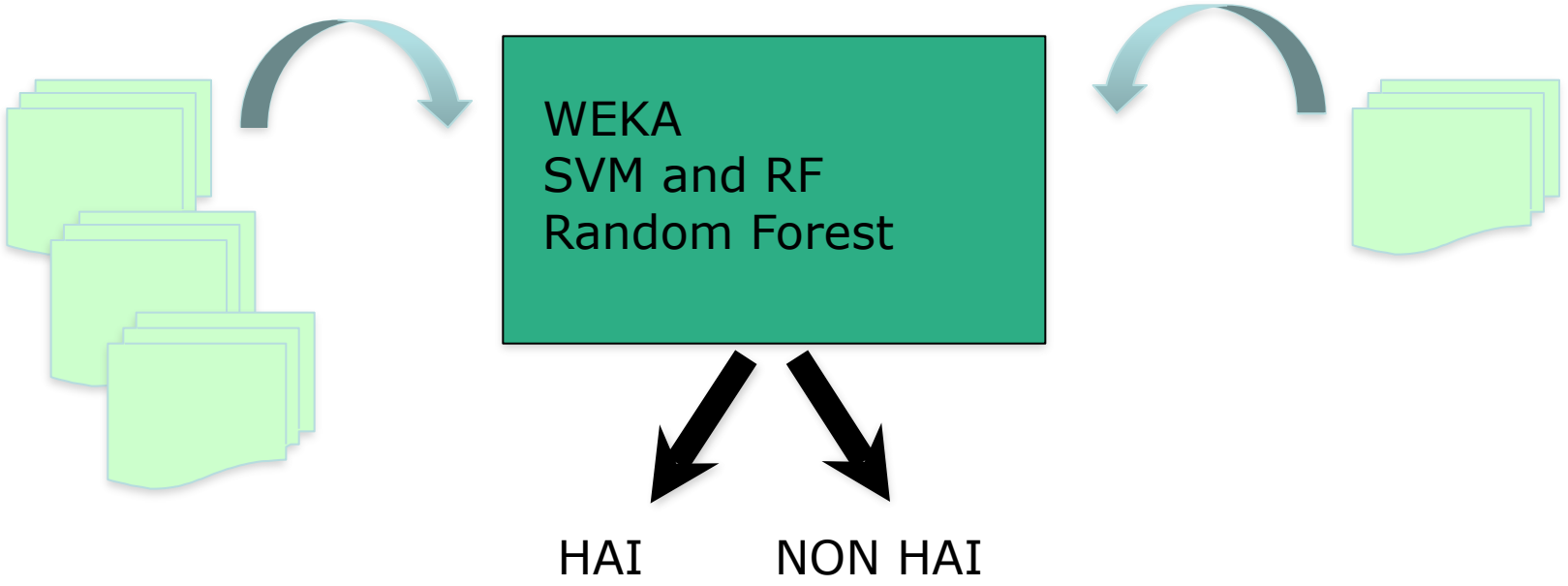
IST infection specific terms
1,045 terminology entries

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

Hospitalisation records for training

Machine Learning

Hospitalisation records for decision



Results detecting HAI

- SVM, Support Vector Machine algorithm 74% recall and 86% precision using Terms + negation
- RF, Random forest, 87% recall and 83% precision, using lemmas
  - See Ehrentraut et al 2014.

Stockholm University

Table 3: Classification results (in percent) of the RF and SVM classifier in combination with the respective preprocessing methods. The best result of each classifier is highlighted.

	RF				SVM			
Preprocessing	Precision	Recall	F_1-score	AUC	Precision	Recall	F_1-score	AUC
Plain	80	86	83	0.84	80	71	75	0.71
NoStopwords	79	87	83	0.84	79	70	74	0.70
Lemma	**83**	**87**	**85**	**0.85**	79	71	75	0.70
IST	80	86	83	0.86	84	74	79	0.76
NegationTagged	80	89	84	0.84	78	71	74	0.70
NegationTagged + IST	81	86	84	0.86	83	73	79	0.75
NegationRemoved	80	87	83	0.84	79	70	73	0.69
NegationRemoved + IST	80	85	82	0.87	**86**	**74**	**80**	**0.77**
TF-IDF 50	74	79	77	0.80	72	65	68	0.63
Tagged	79	84	82	0.85	82	69	75	0.72

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
  - (87% recall and 83% precision for ML-RF 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 unstructured 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 Health care Associated Infections

Research projects

- **MINECAN** - Data and text mining of cancer symptoms and comorbidities in electronic patient records in the Nordic languages, funded The Nordic Information for Action e-Science Center of Excellence, 2014-2019.
- **DADEL** - High-Performance Data Mining for Drug Effect Detection, 2013-2016.
- **AVID** - Avidentifiering för sekundär användning av patientjournaler, 2016.
- **Detect-HAI** - Detection of Healthcare Associated Infections (finalized)

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