Mia Kvist & Kelly Smith

Automated Translation of Medical Records into General Swedish

Dept of Computer and System Sciences, Stockholm University
Health Informatic Centre (HIC), Karolinska Institutet
Clinical text mining group
DSV
Information extraction medical records

- Medical knowledge extraction
- Deidentification (automatic)
- NLP Tools:
  - entity recognition, negation/speculation
- Decision support (alarms, automated coding, etc)
- Surveillance of advert events (Drugs, HAI)
- Translation/simplification
Stockholm EPR Corpus

- Karolinska University Hospital
- 2 milj de-identified patient records
- > 600 000 patients
- 2006-2010
- Various medical, surgical specialities
  + childrens hospital
- No psychiatry
**Today's topics**

On-line records and the need for simple text
Language in health records

Project Text simplification:
- User studies
- Abbreviations
- SCAN – Swedish Clinical Abbreviation Normalizer

Characterization of clinical text (Kelly)
Consumer-centered health communication

- Paradigm shift in health politics
- On-line health records
- Patient empowerment
Hypertoni o DM, brtsm och St lyft anteriort, PCI mot ockluderad proximal LAD med 2 stent. Cirk stabil.
Example record (Anonymized manually)

123 H - ICU 2008-08-21 10:54 1944 Woman

Assessment

Reading our own health records

25% of adult Swedes cannot read texts rich in facts or from authorities

20% leave 9th grade without basic readability skills
Consumer-centered health communication

- Paradigm shift in health politics
- On-line health records
- Information extraction for patients/relatives
- More laymen vocabulary needed in terminologies
Text simplification of medical records

Aim:
Understand what we read in the free text in our own health records,
follow the health care process.

• Patient portals
• PHR - Personal health records
Garner M, Ning Z, and Francis J. Health Expectations 2012;15:283
Health literacy

I Readability

II Comprehensibility

III Communicative effectiveness

Constellation of informational events

Garner M, Ning Z, and Francis J. Health Expectations 2012;15:283
The health divide
Risk to increase the health divide
Rökning dödar.

SMOKING CAUSES MOUTH AND THROAT CANCER

The toxic chemicals in tobacco damage your blood vessels, body's cells and attack your mouth.
Patients’ experiences when accessing their on-line electronic patient records in primary care
C Pyper, J Amery, M Watson and C Crook

Found consultation details easy to understand 80%
Found medication details easy to understand 61%
Found referrals section easy to understand 41%
Found record easy to understand overall 73%
Humerusfraktur

Fracture of humerus

Broken bone in upper arm

Broken arm
Cancer cannot be excluded.
Vilka words are hard to understand?
Centre for easy-reading

- Unusual words
- Oldfashioned words
- Foreign words
- Technical terms
- Jargon
- Compound words
Language in clinical text

Regular Swedish

Medical terminology

Medical jargon
Language in clinical text

Regular Swedish
Missing subjects, passive tens
Misspellings, abbeviations

Medical terminology

Medical jargon
Swedish is a compounding language

The number of terms is endless

Strålbehandlingsplaneringsdatortomografi
Radiationtreatmentplanningcomputertomography

Nefrouretärcystoprostatovesikulektomi
Nephrourethercystoprostatovesiculotomia
Figure 1 - Complementary vocabulary. Ischemic diagnoses are negated as “ischemia” but named as disorders when affirmed.
Negations

No sure clinical signs of appendicitis.
Lung embolus cannot be excluded.
…not only headache but also vomiting.
Language in clinical text

Regular Swedish
Missing subjects, passive tens
Misspellings, abbreviations

Medical terminology
Latin/Greek/English

Medical jargon
Language in clinical text

Foreign words incorporated into Swedish
- Swedish inflections
- mixed spellings

Bronchitis - Latin & English
Bronkit - swedification

Bronchit - misspelling
## Cholecystitis -> kolecystit

<table>
<thead>
<tr>
<th></th>
<th>DAY</th>
<th>X-ray</th>
</tr>
</thead>
<tbody>
<tr>
<td>kolecystit</td>
<td>48</td>
<td>84</td>
</tr>
<tr>
<td>colecystit</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>cholecystit</td>
<td>88</td>
<td>1613</td>
</tr>
</tbody>
</table>

DAY – 0,1 M daily notes, 4 M words
RTG – 0,43 M X-ray reports, 20 M words
<table>
<thead>
<tr>
<th></th>
<th>DAY</th>
<th>RTG</th>
</tr>
</thead>
<tbody>
<tr>
<td>esofagus</td>
<td>176</td>
<td>1339</td>
</tr>
<tr>
<td>oesofagus</td>
<td>4</td>
<td>35</td>
</tr>
<tr>
<td>oesophagus</td>
<td>7</td>
<td>77</td>
</tr>
<tr>
<td>esophagus</td>
<td>21</td>
<td>243</td>
</tr>
</tbody>
</table>
Language in clinical text

Regular Swedish
Missing subjects, passive verbs
Misspellings, abbreviations

Medical terminology
Latin/Greek/English Abbreviations

Medical jargon
Abbreviations X-ray

18 of the top 100 words were abbreviated
• 7 common abbreviations
  \((tel = \text{telephone}, \ cm = \text{centimeter})\),
• 10 domain specific
  \((iv = \text{intravenous})\)
• 1 ambiguous
  \((ca = \text{cancer or circa})\).
Language in clinical text

Regular Swedish
Missing subjects, passive verbs
Misspellings, abbreviations

Medical terminology
Latin/Greek/English Abbreviations

Medical jargon
Neologisms
Language in clinical text

Medical jargon

Making events into actions

The patient has an infarction.
The patient is infarcting.

Consequence:
- Noun to verb
- Word not in terminologies
- The patient as an active agent
Medical jargon
Neo-terminology

Septicemic pat, unclear origin, roundcultured before Zinacef.

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

Function:
an economical, effective and exact tool for daily work
Who do we write the record for?

Understandability depends on whom the intended receiver of the text is.
Who do you write the record for?
Vem skriver du journalen för?

For yourself – as memory notes?
För dig själv – som minnesanteckning?

For colleagues or the health team?
För kollegorna eller vårdlaget?

For the National Board or other authorities?
För Socialstyrelsen el annan myndighet?

For the patient?
För patienten?

Very little
Very much
Automated translation of radiology reports to general Swedish – part of the democratization of health care

Mia Kvist
Martin Duneld
Sumithra Velupillai

Kelly Smith
Lisa Tengstrand

Center for Easy reading
(Centrum för lättläst)
Automated translation of radiology reports to general Swedish

Translation/text simplification/text generation

Aim:
• Machine translation of referrals to radiologists and radiology reports
• Complement to original text
• Not change the content of the report
• Understandable for everybody?
DT thorax-buk: ingen pneumo- eller haemothorax.

CT thorax-abd: no pneumo- or haemothorax.

Datortomografi av bröstkorg-mage visar ingen luft eller blod i lungsäcken.

Computertomography of chest-abdomen shows no air or blood in the pleura ("lungsack").
Resources

Stockholm EPR X-ray corpus

- Year 2009-2010
- General, thoracic, neuro, pediatric radiology
- 434,427 radiology reports questions and answers
- 152,170 unique patients
- All ages: 0-108 år
  40-70 år = 52%
  <18 år = 19%
Personal Health Records in Sweden: Functions preferred by patients

Omran Ibrahim
Masterstudent

Sumithra Velupillai, Maria Kvist
Inst. för data- och systemvetenskap
Overview
• Timelines (diagnoses, drugs)

Help understanding content och test results
• Fact boxes
• Search function
• Regular Swedish language
• Lexicon for medical terminology
• Pictures, curves
• Other languages? Not important

Communication
• Mail
• Ad information
• Chat? Video? Not so eager
USER FRIENDLY

INTERACTIV

Overview

Possible to go deeper into details at choice
Patients interpreting the medical language of discharge summaries

Kirsi Aantaa, Camilla Wide, Maria Kvist, and Sanna Salanterä

Nordiska språk, Åbo Universitet
Karolinska Universitetssjukhuset
Patients interpreting the medical language of discharge summaries

Patients have problems with

- medical terminology
- Abbreviations
- Test results
- Deeper understanding of the meaning of diagnoses and medical events
Patients interpreting the medical language of discharge summaries

Grammar/Syntax

• Incomplete sentences
• Missing subjects

No problem
Survey

Simplification of Radiology reports
Centre for easy reading

Words and sentences
manually simplified step by step

n=100, laymen
- Femurfraktur: Femur fracture
- lårbensbrott: Thighlegbroken
- benbrott i låret: Bonebroken in thigh
- benet är brutet i låret: Leg is broken in thigh
Pleuravätska

lungsäcksvätska

vätska i lungsäcken

vätska runt omkring lungan

Pleural effusion

Lungsackfluid

Fluid in the lungsack

Fluids around the lung
DT skalle undersökningen är utförd utan intravenös kontrast.

Datortomografi skalle undersökningen är gjord utan insprutat i blodkärl kontrastmedel.

Skiktröntgen av huvud är gjord utan att spruta in kontrastmedel.

Vi gjorde skiktröntgen av huvud utan att använda kontrastmedel.

Vi röntgade huvudet.
CT skull examination is performed without intravenous contrast.

Computer tomography scull is done without injecting into bloodvessel contrast.

Layered X-ray of head is done without injecting contrast.

We did layered X-ray of he head without using contrast.

We X-rayed the head.

DT skalle undersökningen är utförd utan intravenös kontrast.

Datortomografi skalle undersökningen är gjord utan insprutat i blodkärl kontrastmedel.

Skiktröntgen av huvud är gjord utan att spruta in kontrastmedel.

Vi gjorde skiktröntgen av huvud utan att använda kontrastmedel.

Vi röntgade huvudet.
De parenkymatösa organen i buken är utan fokala förändringar, liksom mesenterium och tarmar.

De solida organen i buken är utan avgränsade förändringar, liksom tarmkäx och tarmar.

De fasta organen i magen har inga avgränsade förändringar, inte heller tarmarna.

Inga organ i magen har ovanliga förändringar.

No organs in the abdomen has any unusual changes.
Survey

Radiology reports

Centre for easy reading

n=100, laymen

- People will rather have comprehensible than detailed text.
- Simplifications will often, but not always, give a loss of precision and/or information.
Text simplification of medical records

Requires various simultaneous approaches

• lexical exchange for terminology
• abbreviation expansion
• compound splitting
• syntactic simplification
Centrum för lättläst (Centre for easy reading) asserts that radiology reports are too difficult for laymen.

Top 100 lists: All words 49 % are difficult

- Nouns 46 %
- Verbs 47 %
- Adjectives 33 %
Number of words, bigrams, trigrams and sentences in the Stockholm EPR X-ray corpus

<table>
<thead>
<tr>
<th></th>
<th>All words</th>
<th>bigram</th>
<th>trigram</th>
<th>sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Corpus</strong></td>
<td>200703</td>
<td>2534969</td>
<td>5357542</td>
<td>1874464</td>
</tr>
<tr>
<td>types</td>
<td>20290064</td>
<td>17728463</td>
<td>15276077</td>
<td>2567035</td>
</tr>
<tr>
<td>tokens</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Top100</strong></td>
<td>7150511</td>
<td>2759515</td>
<td>1403104</td>
<td>201074</td>
</tr>
<tr>
<td>tokens %</td>
<td>35%</td>
<td>16%</td>
<td>9,2%</td>
<td>7,8%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Number of nouns, verbs and adjectives in the Stockholm EPR X-ray corpus, lemmatized

<table>
<thead>
<tr>
<th></th>
<th>Noun</th>
<th>Verb</th>
<th>Adjective</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corpus:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>types</td>
<td>111468</td>
<td>20351</td>
<td>25278</td>
</tr>
<tr>
<td>tokens</td>
<td>8254868</td>
<td>2079040</td>
<td>2951736</td>
</tr>
<tr>
<td>top100:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tokens 1-100</td>
<td>4389486</td>
<td>1640789</td>
<td>2189288</td>
</tr>
<tr>
<td>%</td>
<td>53%</td>
<td>79%</td>
<td>74%</td>
</tr>
</tbody>
</table>
Implications for a future text simplification system

- A limited set of words and phrases are recurrent.
- These can be used for exchange to more easy-to-read vocabulary.
- Health care-consumer vocabulary synonym lexicon is required.
Implications for a future text simplification system

- Findings dominated by reports of normal conditions and exclusions.
- A pathological finding is not written in a standardized way. Instead, each situation is described in more varied ways.
- More varied formulations probably convey more details that affect the patient directly.
Abbreviations in Swedish Clinical Text - use by three professions

Elin Lövestam\textsuperscript{a}, Sumithra Velupillai\textsuperscript{b} and Maria Kvist\textsuperscript{b, c}

\textsuperscript{a} Dept. of Food, Nutrition and Dietetics, Uppsala University, Uppsala, Sweden

\textsuperscript{b} Dept. of Computer and Systems Sciences, Stockholm University, Stockholm, Sweden

\textsuperscript{c} Dept. of Learning, Informatics, Management and Ethics (LIME), Karolinska Institutet, Sweden
Multiple meanings is manifested for many of the 40 abbreviations for each of the three subsets
Table 3. Frequencies of abbreviations for different expressions of the disorder diabetes mellitus. Abbreviations from the original dietetic corpus were matched against DIET, NURSE AND X-RAY. Note that the subsets are of different sizes and that the different frequencies therefore are not directly comparable.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Original expression</th>
<th>DIET</th>
<th>NURSE</th>
<th>X-RAY</th>
</tr>
</thead>
<tbody>
<tr>
<td>diab</td>
<td>diabetes</td>
<td>118</td>
<td>965</td>
<td>158</td>
</tr>
<tr>
<td>dm</td>
<td>diabetes mellitus</td>
<td>432</td>
<td>38</td>
<td>290</td>
</tr>
<tr>
<td>DM</td>
<td>Diabetes Mellitus</td>
<td>597</td>
<td>69</td>
<td>4219</td>
</tr>
<tr>
<td>DMT1</td>
<td>Diabetes Mellitus type 1</td>
<td>71</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DMT2</td>
<td>Diabetes Mellitus type 2</td>
<td>39</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IDDM</td>
<td>Insulin Dependent Diabetes</td>
<td>416</td>
<td>0</td>
<td>36</td>
</tr>
</tbody>
</table>

Diabetes Mellitus
SCAN:
Swedish Clinical Abbreviation Normalizer
Further Development and
Adaptation to Radiology

Maria Kvist and Sumithra Velupillai
Niklas Isenius
Pat med bruten hö arm. Rtgjour tar över. EKG för säkerhets skull. I övrigt u a.

Pat with broken ri arm. Radiol on call takes over. ECG in case. Otherwise n r.

SCAN:

• Rule-based approaches ported to Swedish
• Lexicons + heuristics
**SCAN** - Lexicons + heuristics

Evaluation SCAN 1 IDENTIFICATION

- 300 medical records
- 2,050 abbreviations
  - annotated by one physician

<table>
<thead>
<tr>
<th>Version</th>
<th>Recall %</th>
<th>Precision %</th>
<th>F-measure %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>87</td>
<td>36</td>
<td>51</td>
</tr>
<tr>
<td>BestListLen3</td>
<td>68</td>
<td>82</td>
<td>75</td>
</tr>
<tr>
<td>BestListLen4</td>
<td>76</td>
<td>81</td>
<td><strong>79</strong></td>
</tr>
<tr>
<td>BestListLen8</td>
<td><strong>83</strong></td>
<td>62</td>
<td>71</td>
</tr>
</tbody>
</table>
## Errors SCAN 1

<table>
<thead>
<tr>
<th>Error type</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Names (people and locations)</td>
<td>54</td>
</tr>
<tr>
<td>Missing terminology in lexicons</td>
<td>21</td>
</tr>
<tr>
<td>Tokenization</td>
<td>19</td>
</tr>
<tr>
<td>Common words</td>
<td>6</td>
</tr>
<tr>
<td>Abbreviation type</td>
<td>ED (%)</td>
</tr>
<tr>
<td>-------------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>Abbreviations, total</td>
<td>11</td>
</tr>
<tr>
<td>Of these abbreviations:</td>
<td></td>
</tr>
<tr>
<td>Acronyms</td>
<td>37</td>
</tr>
<tr>
<td>Shortened words or contractions</td>
<td>63</td>
</tr>
<tr>
<td>Compounds with abbreviation</td>
<td>12</td>
</tr>
</tbody>
</table>

Averages of 3 datasets with 10,000 words in each set.
Compound words

very common that both parts of the word were abbreviated

- *jnlant* - journalanteckning
  engl: record note

- *tablbeh* - tablett behandling
  engl: tablet treatment
Performance of SCAN on two different sublanguages of clinical text.
A majority of the correct expansions are present in the lexicons
Questions?