Exploitation of Structured Knowledge Sources for Question Answering: Future Aspects

Stefan Schulz
Markus Kreuzthaler
Ulrich Andersen
Find non-ontological predications outside SNOMED CT

Approach A

- **SNOMED CT ontological content**
- **SNOMED CT pre-coordinations**
- **SNOMED CT terminological content**

Non-ontological, factoid knowledge learnt from external sources

Examples:
- Insulin is a common treatment for diabetes type 1
- Retinopathy is a typical late complication of diabetes mellitus
- Diabetes type 2 can be controlled by diet

Contingent, probabilistic, default (typically true in certain contexts)

- Examples: 
  - Modifiers like "diabetic" or "terminal" indicate aetiology or severity of a disease
  - The fact that there is a pre-coordinated concept "diabetic nephropathy" but not "diabetic hepatopathy" may indicate that diabetes affects the kidney but not the liver

Result of interpretation of term meanings

Relevance of association (which justifies pre-coordinated concepts)

Universal truths (what holds for all instances of a concept without exceptions)

Confidence

Examples:
- Diabetes mellitus is a disorder of the endocrine system
- Insulin is a peptide
- Metformin contains nitrogen
### Possible predicates between SNOMED semantic types

<table>
<thead>
<tr>
<th></th>
<th>Disease</th>
<th>Finding</th>
<th>Substance</th>
<th>Organism</th>
<th>Body Part</th>
<th>Procedure</th>
<th>Device</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Disease</strong></td>
<td>complications</td>
<td>has sign or symptom</td>
<td>is caused by, is treated by, is prevented by, has metabolite</td>
<td>is caused by, exhibits</td>
<td>is located in</td>
<td>is treated by, is caused by</td>
<td>is caused by</td>
</tr>
<tr>
<td><strong>Finding</strong></td>
<td>sign or symptom of</td>
<td>accompanied by</td>
<td>is caused by</td>
<td>is caused by</td>
<td>is located in</td>
<td>is targeted by</td>
<td>is caused by</td>
</tr>
<tr>
<td><strong>Substance</strong></td>
<td>causes, treats, prevents, is metabolite of</td>
<td></td>
<td>interacts</td>
<td>is affected by</td>
<td>targets</td>
<td>is used by</td>
<td></td>
</tr>
<tr>
<td><strong>Organism</strong></td>
<td>causes, is observed in</td>
<td>causes</td>
<td>is sensitive to</td>
<td>interacts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Body part</strong></td>
<td>is a location of</td>
<td>is a location of</td>
<td>is targeted by</td>
<td>is targeted by</td>
<td></td>
<td>is targeted by</td>
<td></td>
</tr>
<tr>
<td><strong>Procedure</strong></td>
<td>treats, causes</td>
<td>treats, causes</td>
<td>uses</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Device</strong></td>
<td>causes</td>
<td>causes</td>
<td></td>
<td>targets</td>
<td>is used in</td>
<td>is used with</td>
<td></td>
</tr>
</tbody>
</table>

- most of them are non-ontological and therefore not asserted in SNOMED CT
- Knowledge source to be explored: **UMLS co-occurrence matrix**
Example MEDLINE MeSH annotations

MEDLINE bibliographic records (> 20,000,000) are manually annotated using MeSH descriptors.
On MEDLINE concept / concept co-occurrences

- The UMLS provides a co-occurrence matrix

MeSH subheadings, which refine the meaning of C1.
E.g.
CO = complicates   PA = pathology
DI = diagnoses     SU = surgery

```plaintext
C0026683|C0001883|5 |CO=4,DI=2,PA=2,SU=2,RA=1
C0026683|C0001948|1 |ET=1,SU=1
C0026683|C0002475|1 |ET=1,PA=1,TH=1
C0026683|C0003392|1 |DT=1
C0026683|C0003466|2 |SU=2,DI=1,EP=1,ET=1,PA=1
C0026683|C0003611|5 |SU=4,DI=2,PA=2,CO=1,RA=1
C0026683|C0003611|10|SU=8,DI=5,PA=4,CO=3,RA=2,US=2,ET=1
C0026683|C0003614|19|SU=11,PA=10,DI=9,ET=6,CO=5,RA=2,US=2
C0026683|C0003614|21|SU=13,DI=9,CO=8,PA=7,RA=6,ET=5,US=2
C0026683|C0003615|4 |SU=4,US=3,PA=2,RA=2,CO=1,DI=1
C0026683|C0003615|5 |SU=3,CO=2,PA=2,RA=2,DI=1,US=1
C0026683|C0003617|41|DI=23,SU=23,PA=17,US=12,CO=10,ET=5,RA=4,EP=1
```
Induction of SPO triples by MeSH subheading analysis

• Principle: define filtering conditions for each predicate type
  – Semantic types of concepts (mapped to SNOMED CT)
  – Co-occurrence values
  – Subheading distribution  
    } checked against thresholds

• Example: Criteria for: \(<C1; \text{is treated by}; C2>:\)
  – C1 is of the SNOMED type \(\text{Disease or Finding}\)
  – C2 is of one of the types \(\text{Substance, Product, Device, Procedure}\)
  – C1 / C2 co-occurrence above threshold
    • log-likelihood > 6.63, corresponds to p<0.01
  – thresholds of subheading rates
    • DT (drug therapy) > 50% or
    • DH (diet therapy) > 50% or
    • TH (therapy) > 50%

• Implemented: Java + Lucene
What can be done to prevent hyperglycemia?
"How can diabetes mellitus be treated?"

1)
(73211009,116700008|treats,243120004)
(Diabetes mellitus (disorder)|treats,Regimes and therapies (regime/therapy))

Question: "Hvad bygger Diabetes behandling på?" "Tree": "(ROOT (VERB bygger bygge)) (dobj (NOUN Diabetes diabetes)) (possd (NOUN behandling behandling)) (pobj (nobj (PRON Hvad hvad)) (ADP på på))))) (punct (X (? ?))))", "hierarchy1": "Clinical Finding" "conceptID": "73211009", "term1": "diabetes mellitus ", "attribute": "116700008|treats", "hierarchy2": "Procedure" "concept2ID": "243120004", "term2": "regimer og behandlinger", "ESICT_EXPRESSION": "(73211009,116700008|treats,243120004)", "trace": "R14", "jasvar": ",", "nejsvar": ",", "forklarendesvar": "Diabetes behandling bygger på følgende"

Trigger: ASSOCIATED WITH
Generated Lucene Query:
IS TREATED BY sidOne:73211009 AND semTypeOne:(disorder OR finding) AND DT:[00000050 TO 00000100] OR TH:[00000050 TO 00000100] OR DH:[00000050 TO 00000100]
Command line:
java -jar esict.jar tmp c:\DataESICT\luceneIndexCocCdataLogLikeFullSubHead ing
sidOne:73211009 AND semTypeOne:(disorder OR finding) AND semTypeTwo:(substance OR product OR_device OR_procedure) AND (DT:[00000050 TO 00000100] OR_TH:[00000050 TO 00000100] OR_DH:[00000050 TO 00000100]) 20

Answers:
Diabetes mellitus (disorder) IS TREATED BY Hypoglycemic agent (substance) log-like: 3.387,74
Diabetes mellitus (disorder) IS TREATED BY Hypoglycemic agent (product) log-like: 3.387,74
Diabetes mellitus (disorder) IS TREATED BY Regular insulin (substance) log-like: 2.420,36
Diabetes mellitus (disorder) IS TREATED BY Insulin (substance) log-like: 2.420,36
Diabetes mellitus (disorder) IS TREATED BY Insulin product (product) log-like: 2.420,36
Diabetes mellitus (disorder) IS TREATED BY Antineoplastic agent (substance) log-like: 164,00
Diabetes mellitus (disorder) IS TREATED BY Antineoplastic agent (product) log-like: 164,00
Diabetes mellitus (disorder) IS TREATED BY Thiazolidinedione (substance) log-like: 157,26
Diabetes mellitus (disorder) IS TREATED BY Thiazolidinedione (product) log-like: 157,26
Diabetes mellitus (disorder) IS TREATED BY Metformin (substance) log-like: 147,59
Diabetes mellitus (disorder) IS TREATED BY Metformin (product) log-like: 147,59
Diabetes mellitus (disorder) IS TREATED BY Glyburide (substance) log-like: 145,51
Diabetes mellitus (disorder) IS TREATED BY Glyburide (product) log-like: 145,51
Diabetes mellitus (disorder) IS TREATED BY Sulfonylurea and its derivatives (substance) log-like: 132,04
Diabetes mellitus (disorder) IS TREATED BY Sulfonylurea (substance) log-like: 132,04
Diabetes mellitus (disorder) IS TREATED BY Sulfonylurea (product) log-like: 132,04
Diabetes mellitus (disorder) IS TREATED BY Subcutaneous injection (procedure) log-like: 79,87
Diabetes mellitus (disorder) IS TREATED BY Long-acting insulin (substance) log-like: 75,22
Diabetes mellitus (disorder) IS TREATED BY Long acting insulin (product) log-like: 75,22
Outlook

• Approach included into ESICT interface soon
• Known limitations
  – UMLS COOC table lacks important information from MEDLINE (document type, non-human, chemicals)
  – Low granularity of MeSH compared to SNOMED CT
  – Cooccurrences not aggregated in the hierarchy
  – No distinction between hypotheses studied and scientific evidence
• Possible future work:
  – Using MEDLINE source data
  – Using text-mined content from abstracts