Semantic Relations of Binary Compounds Annotated with SNOMED CT

DIMITRIOS KOKKINAKIS
Department of Swedish (the Swedish Language Bank) and the Center for Language Technology in Gothenburg
dimitrios.kokkinakis@svenska.gu.se
OVERVIEW

• Problem Setting, Terminology, Motivation
• Background
• Material
• Method
• Results
• Discussion & Future Work
TERMINOLOGY

Compound: a single orthographic word composed of >1 free/unbound morphemes (that can stand alone) unlike bound ones, which occur only as parts of words

- ‘solid’ /‘closed’: 2 or 3 words appear together as one (housewife)
- ‘open’/‘spaced’: words with space between (heart failure)
- ‘hyphenated’: words with space between (ice-cream)

- ‘transparent’ (compositional): meaning can be derived from the meaning of the individual elements (laser printer)
- ‘opaque’ (non-compositional): meaning cannot be derived from the meaning of the individual elements (guinea pig)
BINARY COMPOUNDS

Binary (noun) compounds: 2 words (nouns)

In a typical binary NC *noun1noun2*, *noun2* is the head and *noun1* is a modifier. In general, the meaning of a NC is a specialization of the meaning of its head – the modifier limits the meaning of the head – thus the meaning of the head determines what kinds of things can be done to it: *what it is made of, what is a part of*, etc.

When there are >1 premodifying noun/adjs (n-ary), the sequence has internal structure and requires bracketing – n-ary NCs are structurally ambiguous

```plaintext
[[[small bowel] obstruction] or [small [bowel obstruction]]]
[n1n2]n3 vs. n1[n2n3]
```
MOTIVATION

- characteristic of many types of technical text
- very productive process
- meaning is usually embedded in Noun Compounds
- language understanding systems needs to be able to interpret NCs in order to ascertain sentence meaning
- Knowledge discovery in text collections is a big challenge for the (Text) Mining, and … life sciences is the most interesting domain of application
- NC understanding is an important component of NLP

Information retrieval – query refinement/expansion; indexing
Summarization; Q&A; Machine translation – alignment
Ontology construction – taxonomic relations (part-whole-relations); Knowledge discovery…
MOTIVATION Cont’d

...despite all available knowledge sources the identification of semantic relations remains a non-trivial task – understanding NCs requires the computer to infer the correct semantic relationship between the words that making up the compound:

What is the relation between:

flu virus – influensavirus => virus that causes/... flu
skin virus – hudvirus => virus that infects/... the skin
vaccine virus – vaccinvirus => virus that produces/... vaccine
child virus – barnvirus => virus that kills/... children

Approach: based on the semantic categories of the nouns making up a NC and using verbs or other part-of-speech as an approximation of meaning
BACKGROUND

How many kinds of relations exist? *contradictory views*

- small set of abstract relations that NCs may imply *Levi ’78*
- the semantics of NCs cannot be exhausted by any finite listing of relationships *Downing ’77*

**Goal:** relations that are sufficiently general to cover significant number of NCs but that can be domain specific enough to be useful in analysis

**Solution:** discover paraphrases (e.g. predicates) that make explicit the “hidden” relations between the nouns in a NounNoun compound

Search for contexts that relate the NC as a relative clause containing a wildcard to be filled with a verb
MATERIAL

• The MEDLEX Corpus
  variety of Swedish text documents related to different medical subfields
  80,000 documents ~45 m tokens
• Swedish SNOMED CT®
  Edition of Aug. 2011, approx. 260.000 “original” terms

[...] perkutan koronar~intervention (PCI) är idag den vanligaste reperfusions~metoden vid ST-höjnings~infarkt, och trombocyt~hämning med klopidogrel är standard~terapi vid PCI med stent~implantation …några negativa effekter på blödnings~mönster, ökad endometrie~proliferation eller cell~förändringar har inte setts vid tillägg av testosteron till gestagen~behandling …
METHOD – Corpus-based interpretation

1. SNOMED CT-annotation, identification and segmentation of compounds
   - select binary NCs where both constituents are assigned a SNOMED CT label (i.e. “lexicalised” NCs are excluded)

2. Syntactic analysis – finding the subconstituents in the corpus and using those to ”label” a relation

3. Predicate paraphrasing in order to propose words as designators of the semantic relation(s) of the constituents

4. Interpretation of the underlying semantics
   - interpretation can be highly context dependent
   - some NCs are non-compositional: (vitamin k/k-vitamin)
   - there can be >1 semantic relations encapsulated
METHOD Cont’d

Exclusion criteria
lexicalised compounds: nagelpsoriasis ("psoriasis of nail")
occurrences of binary compounds with annotations on the head or modifier: <migrän>~kohort ("migraine cohort"); tonårs~<abort> ("teenage abortion")

Inclusion criteria
binary compounds where both the head and modifier are annotated with 2 labels/concepts (78.443 or 15.352 unique)
plasma~homocystein ("plasma homocysteine") [substance+substance]
tyreoidea~hormon ("thyroid hormone") [body structure+substance]
dopamin~neuron ("dopamine neurons") [substance+cell]
arsenik~förgiftning ("arsenic poisoning") [substance+disorder]
# DISTRIBUTION OF the LABELS

<table>
<thead>
<tr>
<th>Modifier</th>
<th>Head</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>body structure</td>
<td>body structure</td>
<td>1687</td>
</tr>
<tr>
<td>disorder</td>
<td>body structure</td>
<td>1489</td>
</tr>
<tr>
<td></td>
<td>organism</td>
<td>1067</td>
</tr>
<tr>
<td>finding</td>
<td>body structure</td>
<td>3417</td>
</tr>
<tr>
<td></td>
<td>substance</td>
<td>1149</td>
</tr>
<tr>
<td></td>
<td>qualifier value</td>
<td>1091</td>
</tr>
<tr>
<td></td>
<td>disorder</td>
<td>2245</td>
</tr>
<tr>
<td></td>
<td>observable entity</td>
<td>1231</td>
</tr>
<tr>
<td></td>
<td>finding</td>
<td>1246</td>
</tr>
<tr>
<td>morph. abnormality</td>
<td>body structure</td>
<td>3798</td>
</tr>
<tr>
<td>observable entity</td>
<td>body structure</td>
<td>1284</td>
</tr>
<tr>
<td></td>
<td>substance</td>
<td>2206</td>
</tr>
<tr>
<td>physical object</td>
<td>body structure</td>
<td>1144</td>
</tr>
<tr>
<td></td>
<td>physical object</td>
<td>1033</td>
</tr>
<tr>
<td>procedure</td>
<td>body structure</td>
<td>1923</td>
</tr>
<tr>
<td></td>
<td>substance</td>
<td>1535</td>
</tr>
<tr>
<td>qualifier value</td>
<td>body structure</td>
<td>2295</td>
</tr>
<tr>
<td></td>
<td>substance</td>
<td>1963</td>
</tr>
<tr>
<td></td>
<td>qualifier value</td>
<td>2049</td>
</tr>
<tr>
<td></td>
<td>disorder</td>
<td>1940</td>
</tr>
<tr>
<td></td>
<td>observable entity</td>
<td>1432</td>
</tr>
<tr>
<td></td>
<td>finding</td>
<td>906</td>
</tr>
<tr>
<td></td>
<td>physical object</td>
<td>1060</td>
</tr>
<tr>
<td></td>
<td>procedure</td>
<td>3426</td>
</tr>
</tbody>
</table>

Binary Compounds with >1000 Occurrences
CASE STUDIES...

...thus for each concept category pair examine a portion of the NCs falling under that category pair, paraphrasing the relation between the nouns in that compound and examining if that relation is the same across all compounds under that category pair.

If that relation (list of paraphrasing words) is the same across all selected compounds then that category pair can be recorded as a *rule* linked to the relation produced.

If several relations can be produced for a given category pair, analysis needs to descend the SNOMED CT hierarchies (if possible - not done yet!)
CASE STUDY: +disorder

Body structure+Disorder: hud~mykos, hud~porfyrier, hud~sarkoidos, hud~vaskulit, ...

Disorder+Procedure: allergi~vaccination, bröstcancer~screening, damp~diskussionen, demens~utbildning, difteri~vaccinering, hypotensions~narkos, klamydia~provtagning,...

Disorder+Disorder: allergi~chock, alzheimer~demens, alzheimer~encefalopati, appendicit~abscess, astma~allergi, cirros~ascites, diabetes~dyslipidemi, gonorré~salpingit, sarkoidos~artrit, tuberkulos~pleurit, tuberkulos~spondylit,...

Hypothesis: can NCs pairs with same labels be hypothesized to express the same relation? For certain combinations of classes (category pairs) is it probable that the same relation applies?
**GOAL** Characterize the semantic relationship that occur between the noun constituents by discovering a set of predicates that paraphrase the NC.

Paraphrases guided by e.g. verbal predicates can capture more finer-grained semantic representations.

- N1 is paraphrased as a post-modifying rel. clause

  *e.g. bleedings disease as: disease that can-cause bleeding*

- ...or expressed by a preposition phrase

  *skin cancer as: cancer of/in/... the skin*

Extracted part-of-speech annotated sentences from the corpus containing both N1&N2 and filtered out relevant contexts based on the patterns:

- N2 REL-CLAUSE-INDICATOR ADV? AUX* VERB
  PARTICLE? PREP? ARTICLE? ADJECTIVE(S)* N1

- N2 PREP ARTICLE? N1
RESULTS– Paraphrasing…

muskler/NCUPNI som/PH000 stabiliserar/VMIPA foten/NCUSND
muskler/NCUPNI som/PH000 inte/R0 sitter/VMIPA nära/R0 leder/NCUPNI
muskler/NCUPNI som/PH000 rör/VMIPA extremiteten/NCUSND
muskler/NCUPNI som/PH000 trycker/VMIPA på/S urinblåsan/NCUSND
muskler/NCUPNI som/PH000 omger/VMIPA blockärden/NCUSND
muskler/NCUSNI som/PH000 rör/VMIPA skulderblad/NCUSNI
muskler/NCUSNI som/PH000 stöder/VMIPA ryggraden/NCUSND
muskler/NCUSNI som/PH000 förenar/VMIPA underäknen/NCUSND
muskler/NCUSNI som/PH000 lyfter/VMIPA ögonlocket/NCNSND
muskler/NCUPNI som/PH000 omsluter/VMIPA lufrören/NCUSND
muskeln/NCUSNI som/PH000 rör/VMIPA skulderblad/NCUSNI
muskeln/NCUSNI som/PH000 lymner/VMIPA knäet/NCNSND
muskeln/NCUSNI som/PH000 omger/VMIPA underäknen/NCUSND
muskeln/NCUSNI som/PH000 sitter/VMIPA rast/R0 i/S underäknen/NCUSND
muskeln/NCUSNI som/PH000 omger/VMIPA förmaken/NCUSND…

"surrounds"

"attach"
RESULTS– Paraphrasing…

**Body structure+Disorder**: hud~mykos, hud~porfyrier, hud~sarkoidos, hud~vaskulit, ...
Disorder { ’i’ (in), ’på’ (on) } Body structure
Disorder { ’drabbar’ (affects), ’sitter’ (sit), ’sprider’ (spread) } Body structure

**Organism+Disorder**: listeria~infektion, listeria~meningit, listeria~perikardit, listeria~sepsis, ...
Disorder { ’med’ (with), ’av’ (of), ’bero på’ (depends on) } Organism
Disorder { ’orsakas (caused-by) } Organism

**Qualifier value+Disorder**: recidiv~infektion, recidiv~pneumotorax, recidiv~varicer, recidiv~tonsillit, ...
Disorder { ’med’ (with) } Qualifier value
Disorder { ’är’ (is), ’verkar’ (seem) } Qualifier value
RESULTS– Paraphrasing…

**Body structure+Morphological abnormality**: hud~melanom, hud~atrofi, hud~vårta; hud~defekt, hud~mastocytos, ...

Body structure \{ ‘av (of), ’på’ (on), ’innanför’ (inside) \} Morph. abn.

Body structure \{ ‘förekomma’ (exists), ’upptå’ (appears) \} Morph. abn.


*maskin~kropp

??? \{ - \} Body structure

??? \{ - \} Body structure

**Body structure+Body structure**: bröst~muskulatur, bröst~körtlar; bröst~epitel, bröst~parenkym, bröst~ryggmärg, ...

Body structure \{ ‘vid’ (at), ‘under’, ‘på’ (on) \} Body structure

Body structure \{ ‘får’ (makes), ’skiljer/separera’ (separate), ‘går över’ (covers) \} Body structure
<table>
<thead>
<tr>
<th>Classes</th>
<th>NCs</th>
<th>Paraphrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP+BP</td>
<td>lår<del>muskel rygg</del>muskel tum<del>muskulatur thorax</del>skelett bäcken<del>led knä</del>skelett näs~brosk ...</td>
<td>sitter(11), ligger(9), skyddar(5), börjar(3), stabilisera(2), finns(2), fäster(2), håller_ihop(2), ...</td>
</tr>
<tr>
<td>BP DISFND</td>
<td>muskel<del>värk muskel</del>smärtaskelett<del>smärta knäleds</del>smärta ryggrads<del>inflammation muskel</del>kakexi epiglottis~ödem</td>
<td>drabbar(9), uppkommer (6), beror(5), sitter(5), finns(4), sprider(3), känns(2), kommer(2), får(2) ...</td>
</tr>
<tr>
<td></td>
<td>BP-PRCD</td>
<td>lever<del>biopsi tunntarms</del>biopsi svalg<del>reflex pankreas</del>diagnostik lever<del>palpation gallvägs</del>scint ileum<del>biopsi lever</del>provtagning av; på från stängers av av från från</td>
</tr>
</tbody>
</table>
FUTURE WORK

• naming the relations using hyperonyms from a general resource should be explored in the future for instance the Swedish Swesaurus/WordNet <spraakbanken.gu.se/eng/resource/swesaurus>

• searching the whole WEB, than merely a sampled corpus, for support of a larger set of predicates should be explored (“web as a corpus”)

• more paraphrasing patterns

• more “relaxed” patterns

• groupings

• compare with the existing SNOMED relations
CONCLUSIONS

Apply the juxtaposition of hierarchy membership within SNOMED CT to determine the relations between pairs of nouns in Swedish NCs

Exploring SNOMED CT’s concept labels results that the data becomes less sparse and more feasible to e.g. compute statistics over occurrences of codes than over words

Automatic extraction of semantic relationships between instances of a semantic resource is necessary in order to attach richer semantic metadata to documents than it is currently possible, useful for a range of language technology applications

Predicate paraphrasing was used in order to propose a list of words as designators of the semantic relation of the constituents – a method with varied degree of success depending on the categories examined – for many concept combinations the results were still sparse...