Detection of underspecifications in SNOMED CT concept definitions using language processing

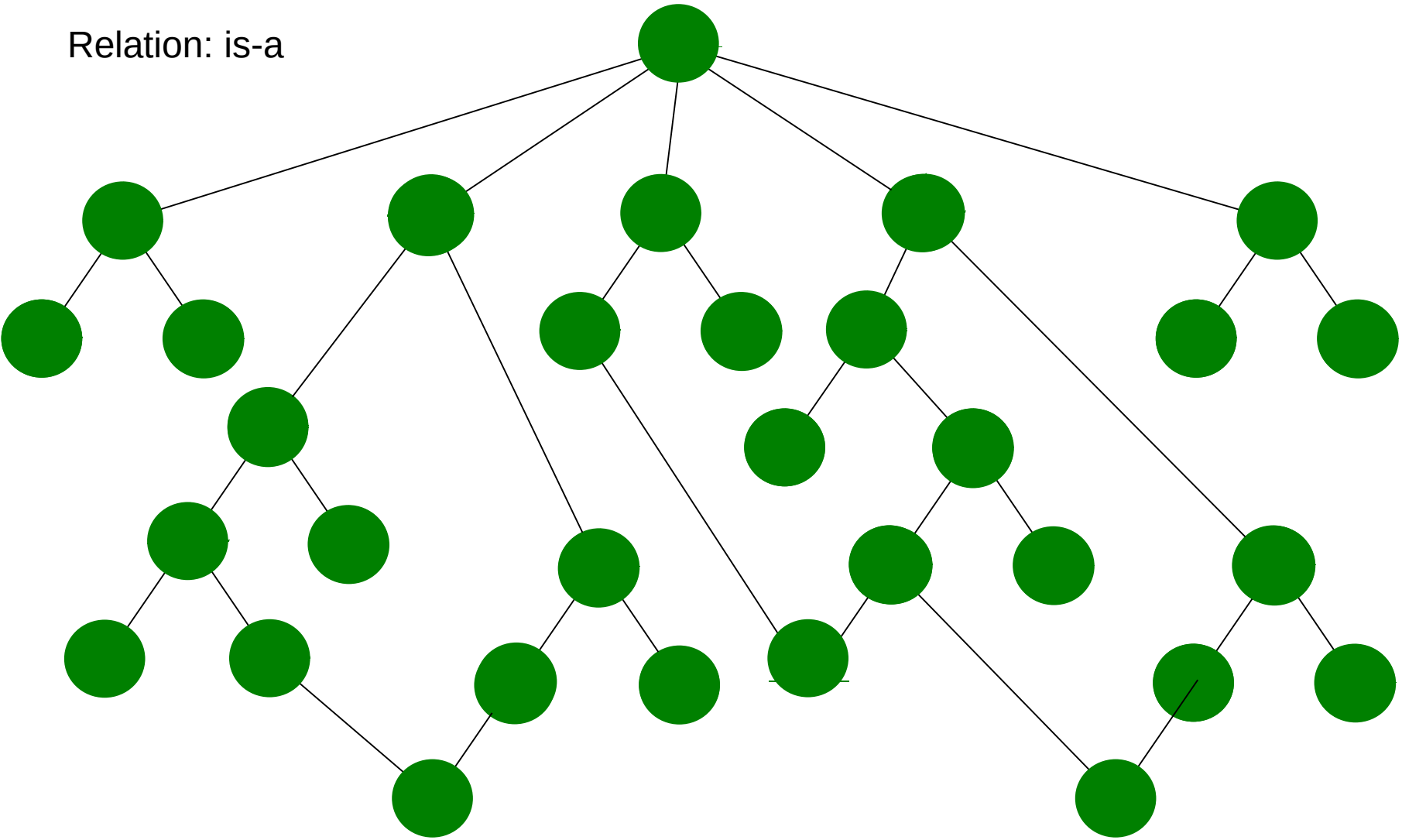
Edson Pacheco[1,2], Holger Stenzhorn[3], Percy Nohama[1], Jan Paetzold[3,4], Stefan Schulz[2,3,4]

[1]Federal Technical University of Paraná (UTFPR), Curitiba, Brazil;
[2]Pontifical Catholic University of Paraná (PUCPR), Curitiba, Brazil;
[3]Institute of Medical Biometry und Medical Informatics, University Medical Center Freiburg;
[4]AVERBIS GmbH, Freiburg, Germany

SNOMED CT

- “Standardized Nomenclature of Medicine - Clinical Terms”
- Comprehensive clinical terminology
  (> 300,000 representational units)
- Concepts are arranged in extensive taxonomic (is-a) hierarchies

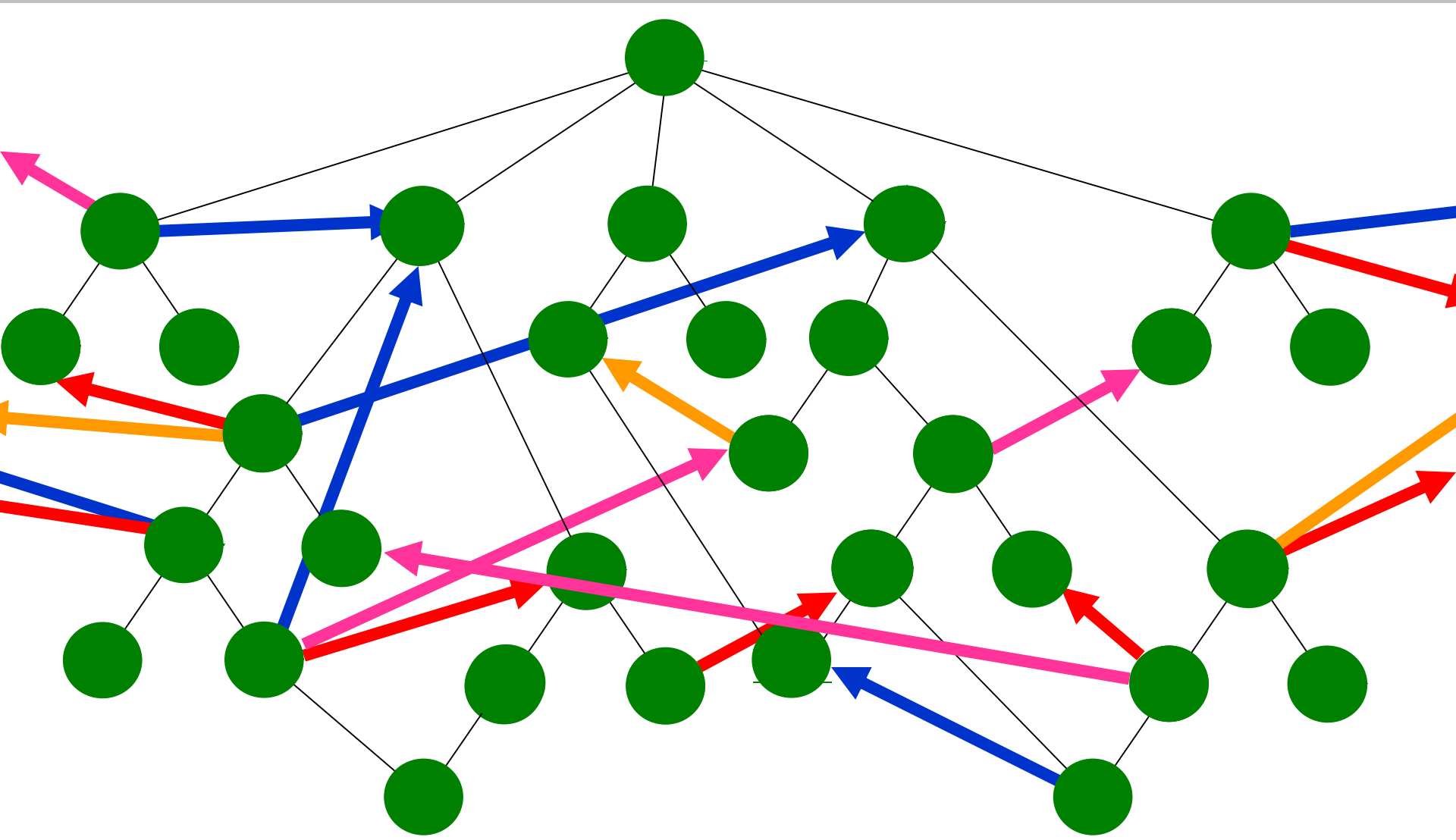
Taxonomic Structure of SNOMED CT



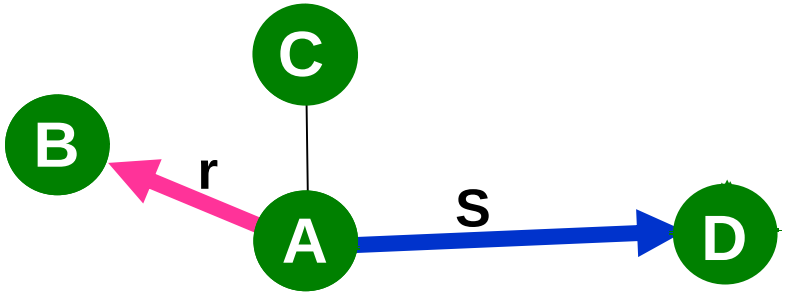
SNOMED CT

- “Standardized Nomenclature of Medicine - Clinical Terms”
- Comprehensive clinical terminology
  (> 300,000 representational units)
- Concepts are arranged in extensive taxonomic (is-a) hierarchies
- Cross-reference between concepts from several branches via semantic relations obeying description logics semantics

Cross-reference between SNOMED CT concepts



SNOMED CT semantics in a nutshell



A: concept
C: taxonomic parent
B, D: attributes

A – isA – C
A – r – B
A – s – D

A subClassOf
C and
r *some B and*
s *some D*

$\mathcal{EL}+$ description logics

SNOMED CT

- “Standardized Nomenclature of Medicine - Clinical Terms”
- Comprehensive clinical terminology
  (> 300,000 representational units)
- Concepts are arranged in extensive taxonomic (is-a) hierarchies
- Cross-reference between concepts from several branches via semantic relations obeying description logics semantics
- Burden of terminology content maintenance and quality assurance

SNOMED CT

- “Standardized Nomenclature of Medicine - Clinical Terms”
- Comprehensive clinical terminology
  (> 300,000 representational units)
- Concepts are arranged in extensive taxonomic (is-a) hierarchies
- Cross-reference between concepts from several branches via semantic relations obeying description logics semantics
- Burden of terminology content maintenance and quality assurance
- To be supported by automated approaches

Looking for underspecifications of cross-linkage

- Nearly half (45.2%) of the SNOMED CT concepts (132,125) have no attributes.
- Textual descriptions suggest composed meanings
- Examples:
  - *Cerebral function*
    - only related to its parent *Nervous system function*
    - expected relation with *Brain structure* missing
  - *Hepatitis notification*
    - only related to its parent *Disease notification*
    - expected relation with *Inflammatory disease of liver* missing

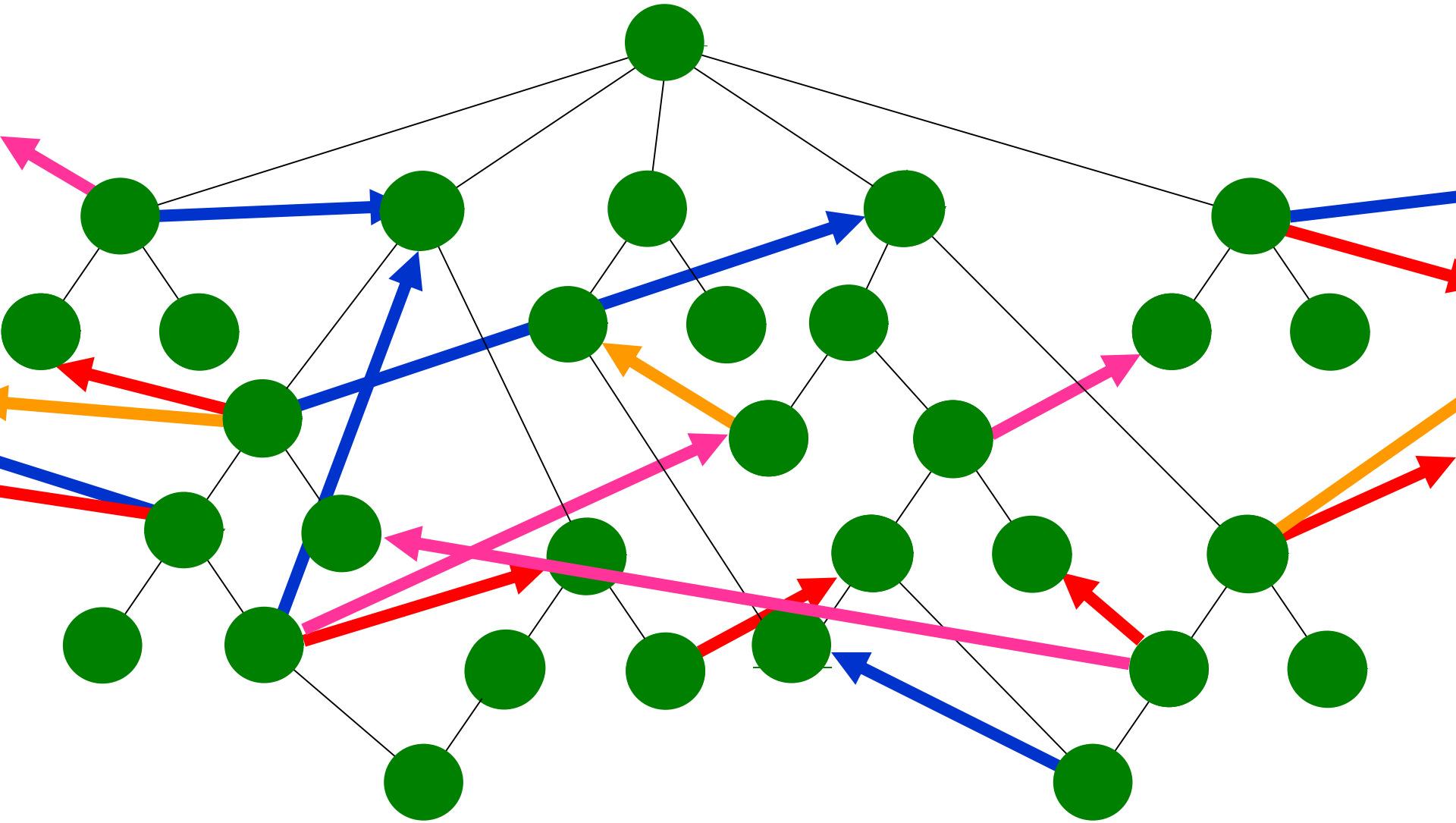
Automatic suggestion of attributes

- Source: 01/2009 release of SNOMED CT

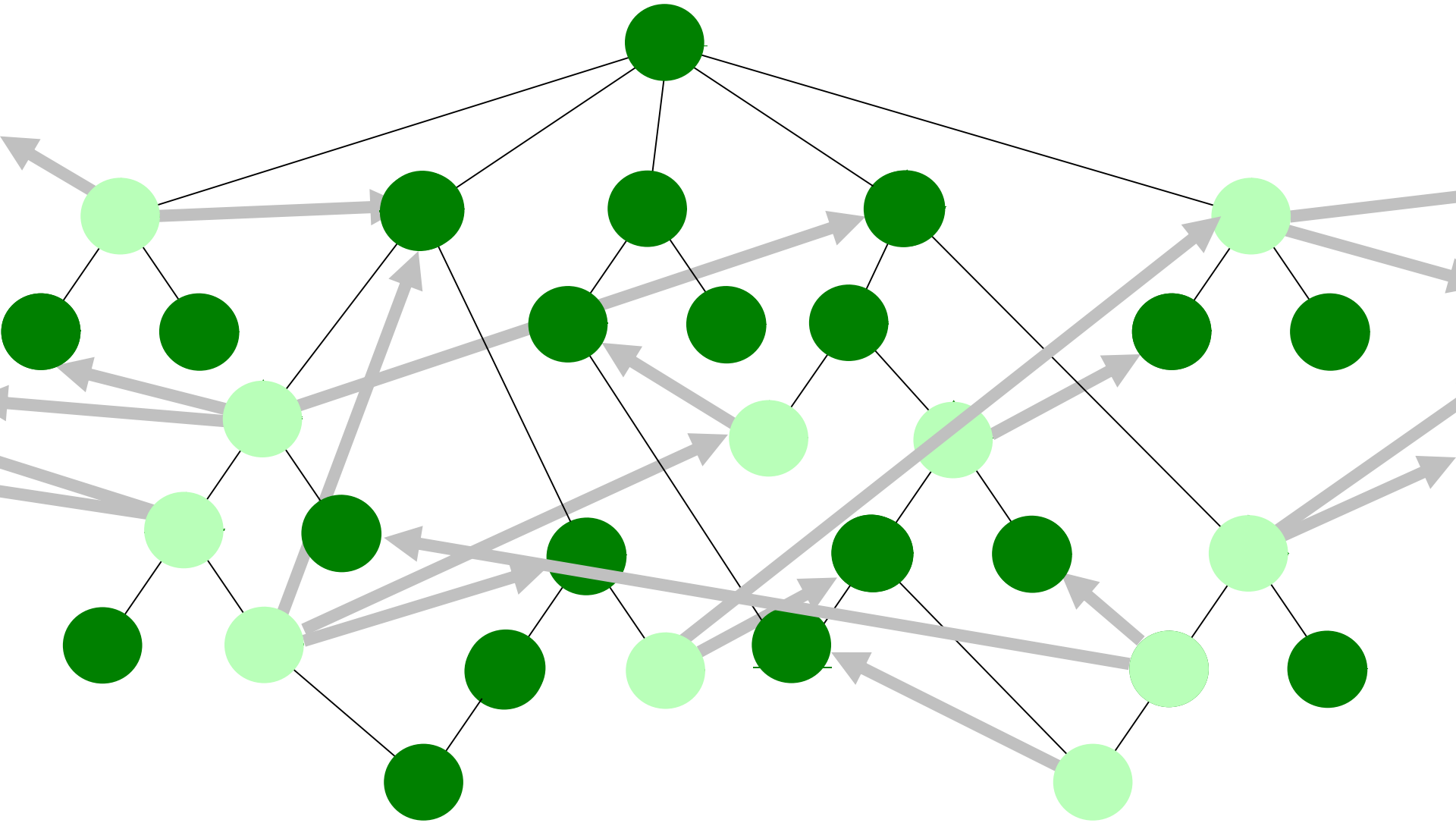
Automatic suggestion of attributes

- Source: 01/2009 release of SNOMED CT
- Algorithm:
  1. Identify non-attributed concepts

Select non-attributed SNOMED CT concepts



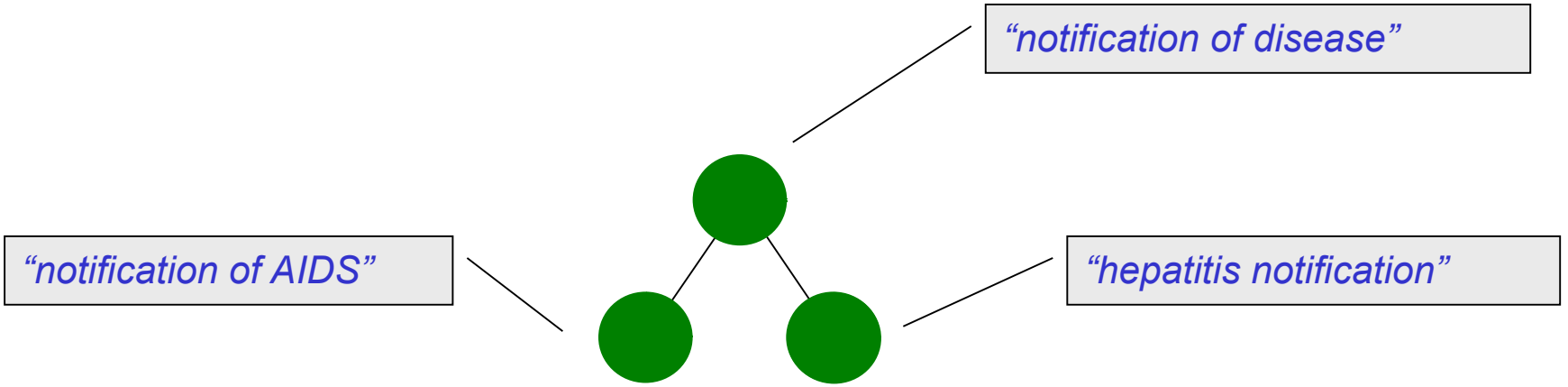
Select non-attributed SNOMED CT concepts



Automatic suggestion of attributes

- Source: 01/2009 release of SNOMED CT
- Algorithm:
  1. Identify non-attributed concepts
  2. Extract concept names

Extract names of non-attributed concepts

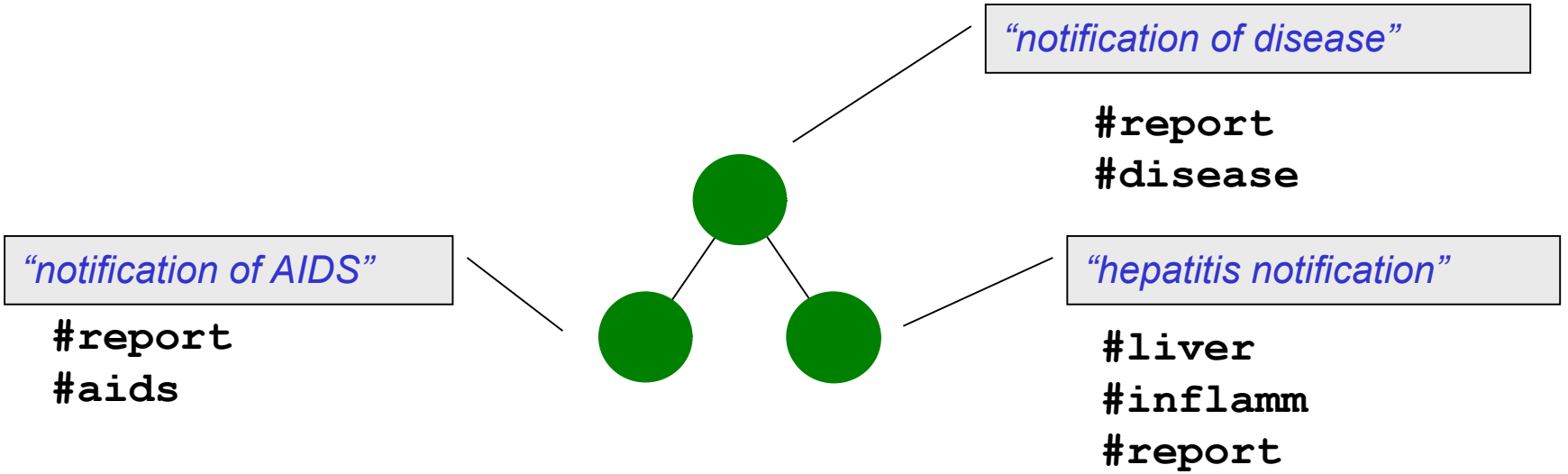


Automatic suggestion of attributes

- Source: 01/2009 release of SNOMED CT
- Algorithm:
  1. Identify non-attributed concepts
  2. Extract concept names
  3. Perform semantic abstraction from word to sets of morphosemantic identifiers (MIDs)

Perform morphosemantic abstraction

Using MorphoSaurus* morphosemantic indexing



*Markó K, Schulz S, Hahn U: MorphoSaurus - Design and Evaluation of an Interlingua-based, Cross-language Document Retrieval Engine for the Medical Domain. Meth Inf Med 4/2005(44): 537-545.

Automatic suggestion of attributes

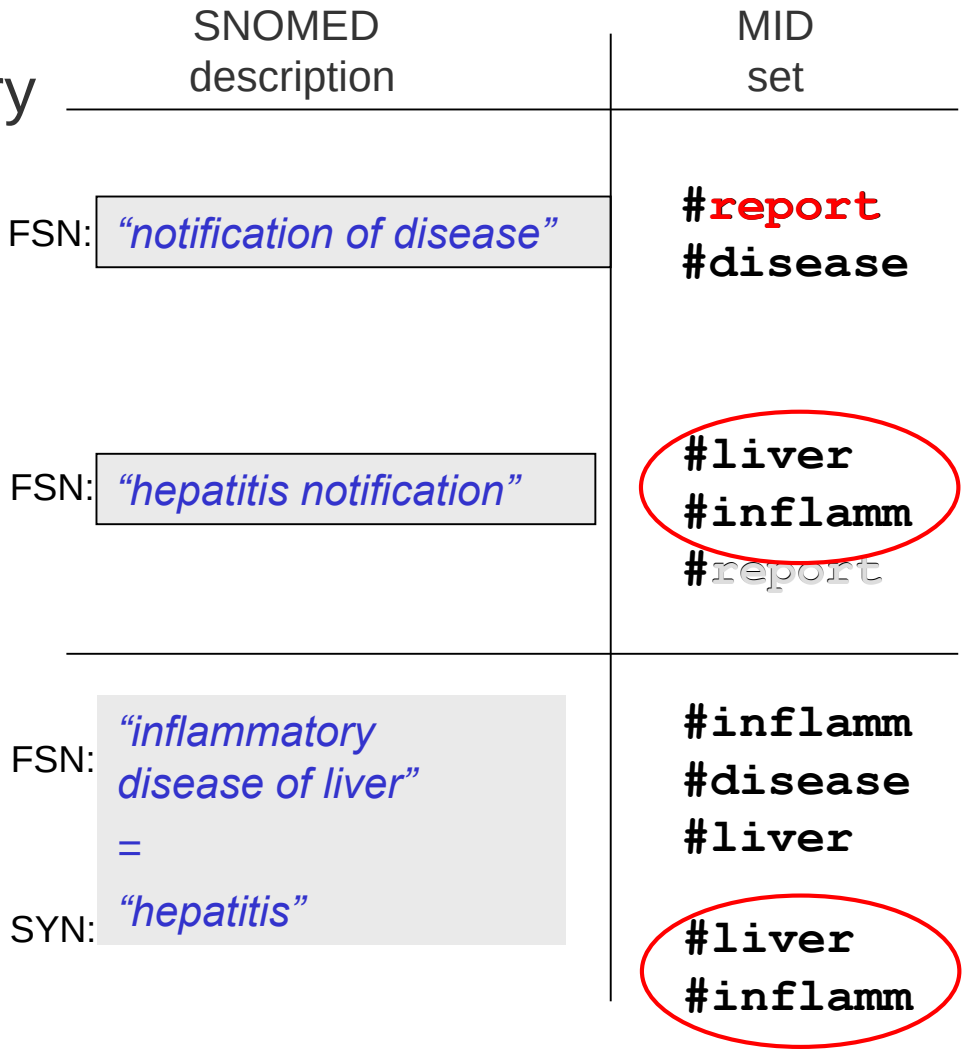
- Source: 01/2009 release of SNOMED CT
- Algorithm:
  1. Identify non-attributed concepts
  2. Extract concept names
  3. Perform semantic abstraction from word to sets of morphosemantic identifiers (MIDs)
  4. Compare MID sets between children and parents and reduce child sets

Automatic suggestion of attributes

- Source: 01/2009 release of SNOMED CT
- Algorithm:
  1. Identify non-attributed concepts
  2. Extract concept names
  3. Perform semantic abstraction from word to sets of morphosemantic identifiers (MIDs)
  4. Compare MID sets between children and parents and reduce child sets
  5. Match reduced child set against MID representations of all SNOMED descriptions

Matching heuristics

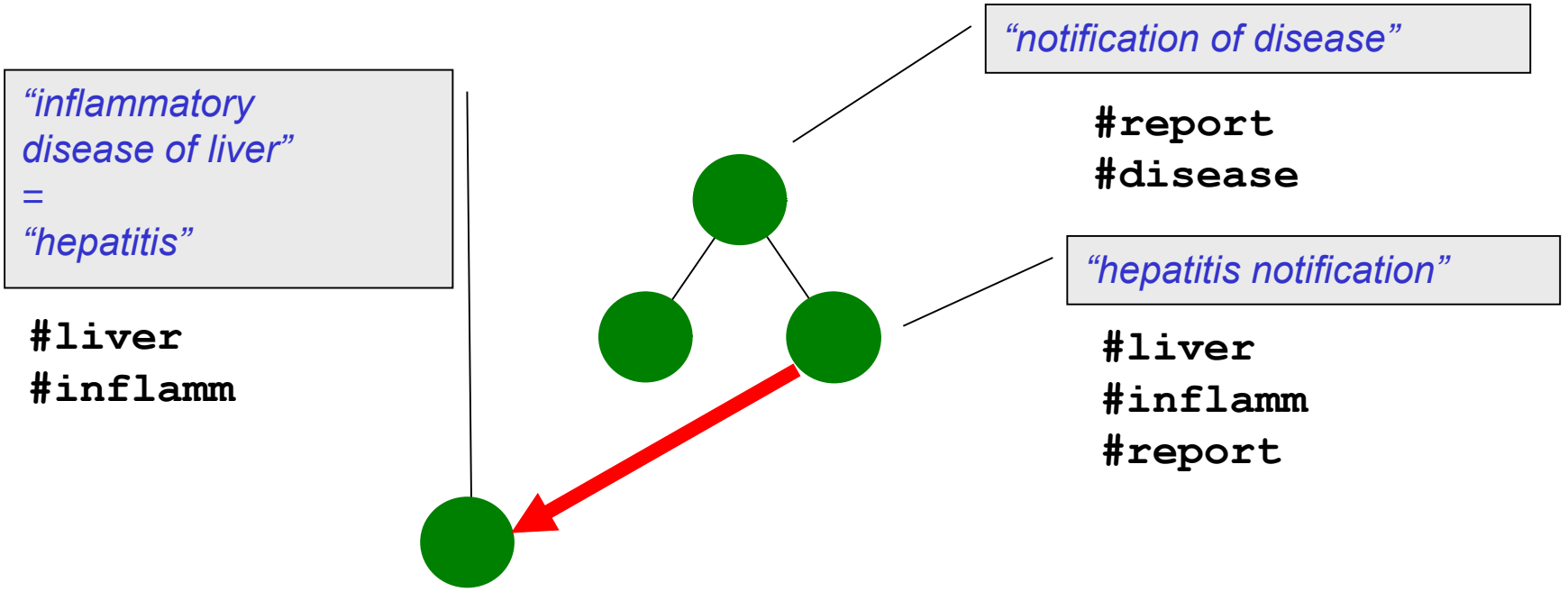
- For the FSN MID set of every non-attributed concept:
  - remove MID that occurs in any of this concept's parents
  - check whether the remainder set coincides with the MID representation of some other SNOMED CT concept, considering all descriptions (FSNs, PTs, synonyms)
  - consider this concept a refinement candidate



Automatic suggestion of attributes

- Source: 01/2009 release of SNOMED CT
- Algorithm:
  1. Identify non-attributed concepts
  2. Extract concept names
  3. Perform semantic abstraction from word to sets of morphosemantic identifiers (MIDs)
  4. Compare MID sets between children and parents and reduce child sets
  5. Match reduced child set against MID representations of all SNOMED descriptions
  6. Suggest candidates for refining attributes

Addition of refinement candidate



*Markó K, Schulz S, Hahn U: MorphoSaurus - Design and Evaluation of an Interlingua-based, Cross-language Document Retrieval Engine for the Medical Domain. Meth Inf Med 4/2005(44): 537-545.

Evaluation Methodology

- For each of 14 SNOMED subhierarchies: random sample of 20 underspecified concepts, compared to attribute refinement candidate proposed by the system
- For each of the sample concept verify
  1. whether this concept should be refined
  2. whether one of the suggested refinement candidates can be plausibly used for refinement.
- Performed by two domain experts. Double rating for interrater agreement measurement: 25%

Results of retrieval experiments

Results of retrieval experiments

SNOMED hierarchies	Active Concepts	Non-attributed Concepts n	Non-attributed Concepts %	Refinement candidates n	Refinement candidates %	Analysis of samples(n=20) refine-ment justified	Analysis of samples(n=20) **correct suggest ion**	Sample based estimation refinable concepts	Sample based estimation with correct suggest-ions
Organism	31840	31840	100.0	4973	15.6	0%	**0%**	**0**	**0**
Substance	23554	23554	100.0	8627	36.6	55%	**35%**	**4700**	**3000**
body structure	25637	22386	87.3	15076	58.8	5%	**0%**	**800**	**0**
qualifier value	8823	8823	100.0	3533	40.0	0%	**0%**	**0**	**0**
observable entity	7885	7885	100.0	3647	46.3	70%	**50%**	**2600**	**1800**
Finding	32780	5356	16.3	2253	6.9	90%	**75%**	**2000**	**1700**
physical object	4408	4408	100.0	1339	30.4	85%	**80%**	**1100**	**1100**
morphologic abnormality	4297	4289	99.8	2164	50.4	80%	**60%**	**1700**	**1300**
Occupation	3843	3843	100.0	1330	34.6	75%	**10%**	**1000**	**100**
Product	19310	3541	18.3	686	3.6	100%	**60%**	**700**	**400**
Event	3578	3529	98.6	447	12.5	85%	**45%**	**400**	**200**
Disorder	63874	2812	4.4	1080	1.7	90%	**60%**	**1000**	**600**
Procedure	47764	2256	4.7	1001	2.1	85%	**65%**	**900**	**700**
Others	14511	7603	52.4	2396	16.5	75%	**60%**	**1800**	**1400**
TOTAL	292104	132125	45.2	48552	16.6			**18700**	**12300**

Results

- Interrater agreement (Kohen's kappa):
  - A concept should be refined: 0.55 (low !)
  - There is a proposed refinement candidate: 0.74
- Estimation: approximately 18,000 SNOMED CT concepts can be refined.
- Problematic suggestions:
  - *Macaroni* for *Macaroni maker*
  - *Canada* for *Salmonella canada*
  - *Acyl carnitine* for *Acylcarnitine hydrolase*
  - *First* for *Female first cousin*
    (already fully defined by the intersection of *First cousin* and *Female cousin*)

Conclusions

- Many SNOMED CT concepts are underdefined and can / should be refined
- The proposed methodology was useful to detect underspecifications
- Large difference between SNOMED hierarchies re harvesting and approval of refinement candidates
- “Grey areas”
  - many proposed refinements are debatable
  - only part of refinement candidates not retrieved due to restrictions of the methodology
- Should be considered for future SNOMED CT editing policies

Thank You!
Contact:

Stefan Schulz

http://purl.org/steschu

Acknowledgements:

CNPq, Brazil:
550830/2005-7

BMBF-IB, Germany:
BRA05/022

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

averbis
medical language technology

UNIVERSITÄTS KLINIKUM
FREIBURG