Ontology mapping needs context

Frank van Harmelen
Vrije Universiteit Amsterdam

vrije Universiteit

introductory comments on ontologies & contexts

Ontologies

- we know what they are
  “consensual, formalised models of a domain”
- we know how to make and maintain them
  (methods, tools, experience)
- we know how to deploy them
  (search, personalisation, data-integration, ...)

Main remaining open questions

- Automatic construction (learning)
- Automatic mapping (integration)

Contexts

?

- I don't really know what they are... Quote from CfP: "Earlier workshops were mostly focused on what contexts and ontologies are".
- At least (?) two views:
  - context as "module", **`ist(p,c)`**
  - context as "relevant knowledge", "contextual meaning"
- I will use 2nd meaning

context-specific nature of knowledge

Opinion poll

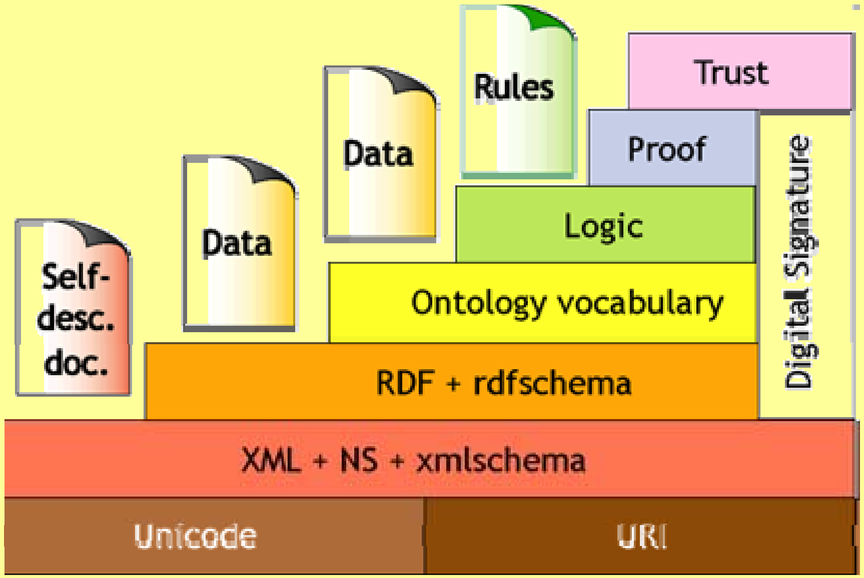
left

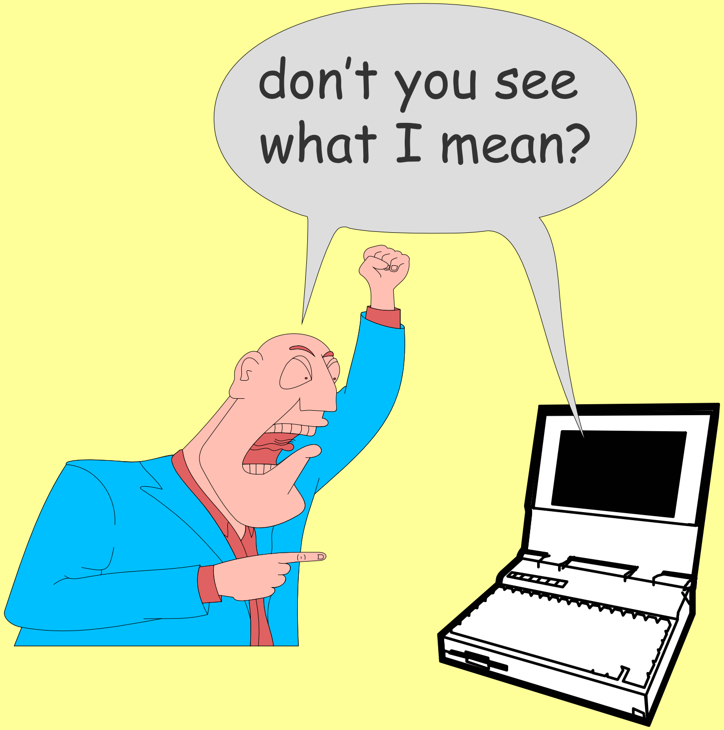
meaning of a sentence
is **only** determined
by the sentence itself,
and **not** influenced by
the surrounding
sentences,
and **not** by the situation
in which the sentence
is used

right

meaning of sentence
is **not only** determined
by the sentence itself,
but is **also** influenced by
by the surrounding
sentences,
and **also** by the situation
in which the sentence
is used

Opinion poll
left right

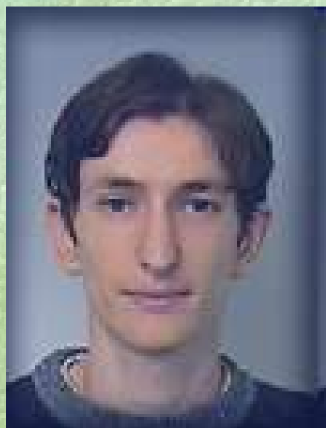




Agenda for talk

- Does this “context dependency” also hold for ontology mapping?

- Intuitively: yes, obviously
- More precisely:
  - can context compensate for lack of structure in source and target?
  - is more context knowledge better?
  - is richer context knowledge better?

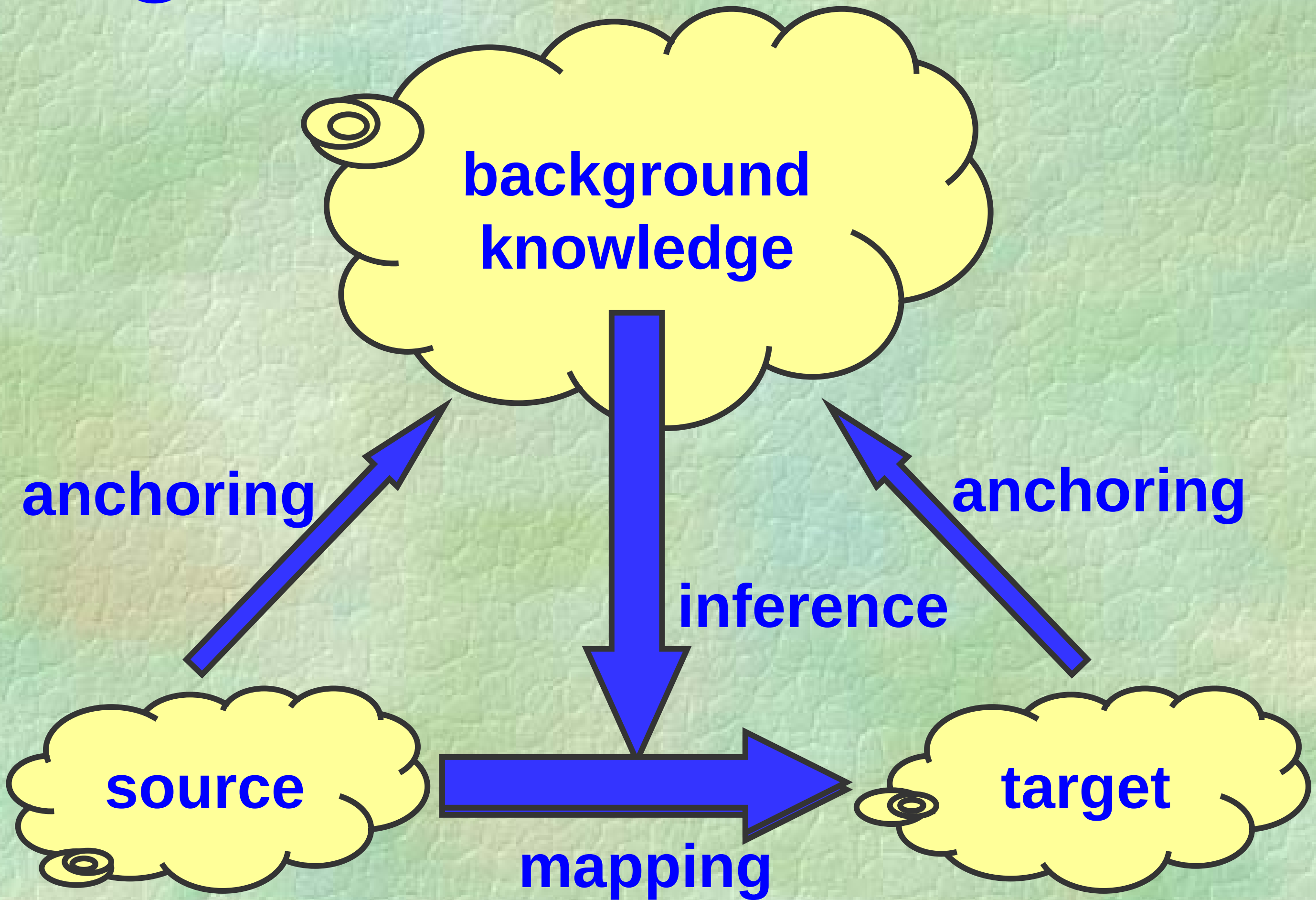


This work with
Zharko Aleksovski &
Michel Klein

vrije Universiteit

Does context knowledge help mapping?

The general idea



source and target vocab's

- OLVG "problem-list":
  - around 3000 problems in a flat list
  - based on ICD9 + "classificatie van verrichtingen"
  - contains general and specific categories
    - implicit hierarchy
    - e.g.6 types of Diabetes Mellitus, many fractures
  - some redundancy because of spelling mistakes
  - used to keep track of the problems of patients during the whole stay at the ICU
- **OLVG-1400**:
  - the subset used in the first 24 hour of stay since 2000 (contains data about 3602 patients)
- **AMC**: similar list, but from different hospital

Context ontology used

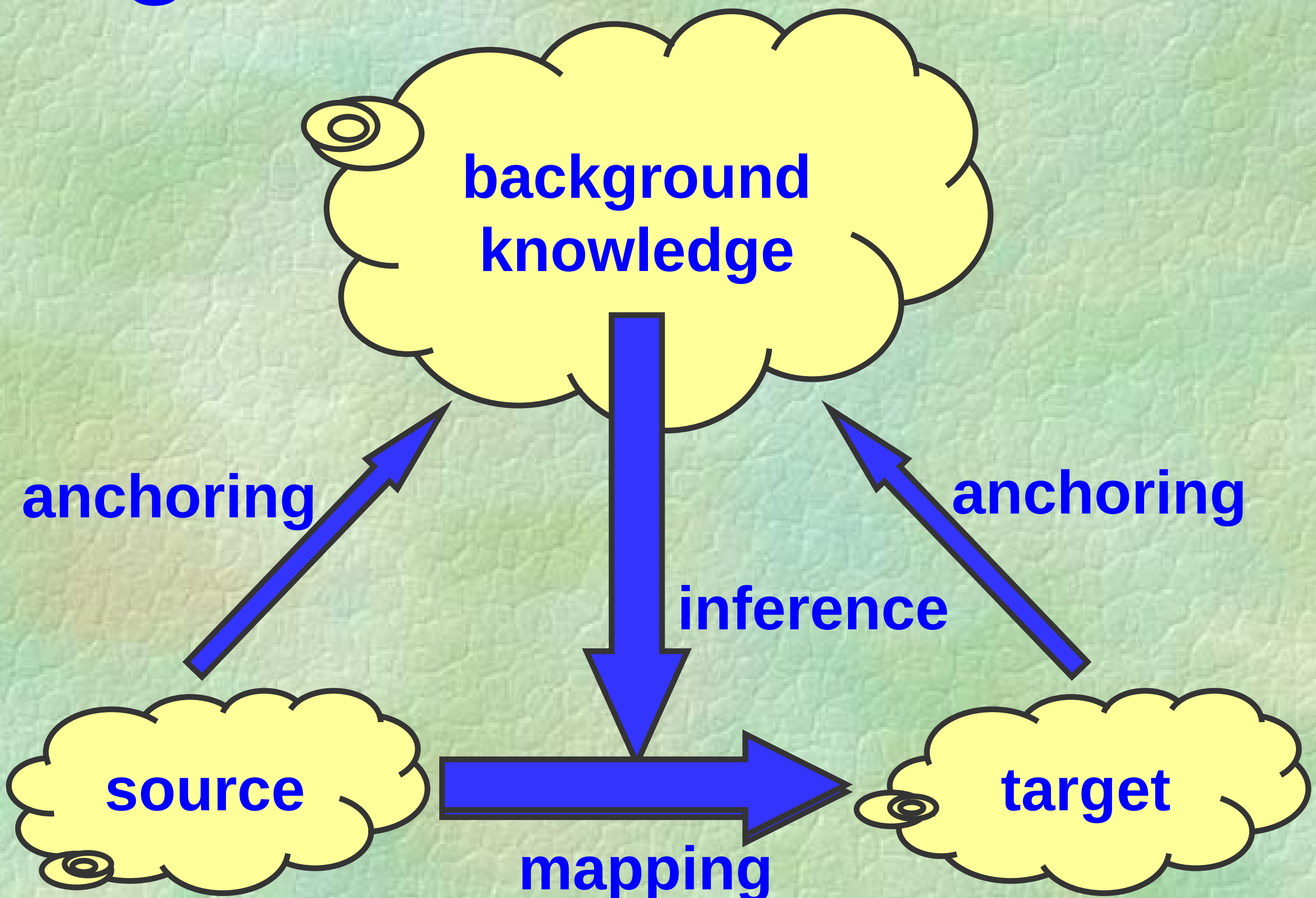
- **DICE**:
  - 2500 concepts (5000 terms), 4500 links
  - Formalised in DL
  - five main categories:
    - tractus (e.g. nervous_system, respiratory_system)
    - aetiology (e.g. virus, poising)
    - abnormality (e.g. fracture, tumor)
    - action (e.g. biopsy, observation, removal)
    - anatomic_location (e.g. lungs, skin)

Baseline: Linguistic methods

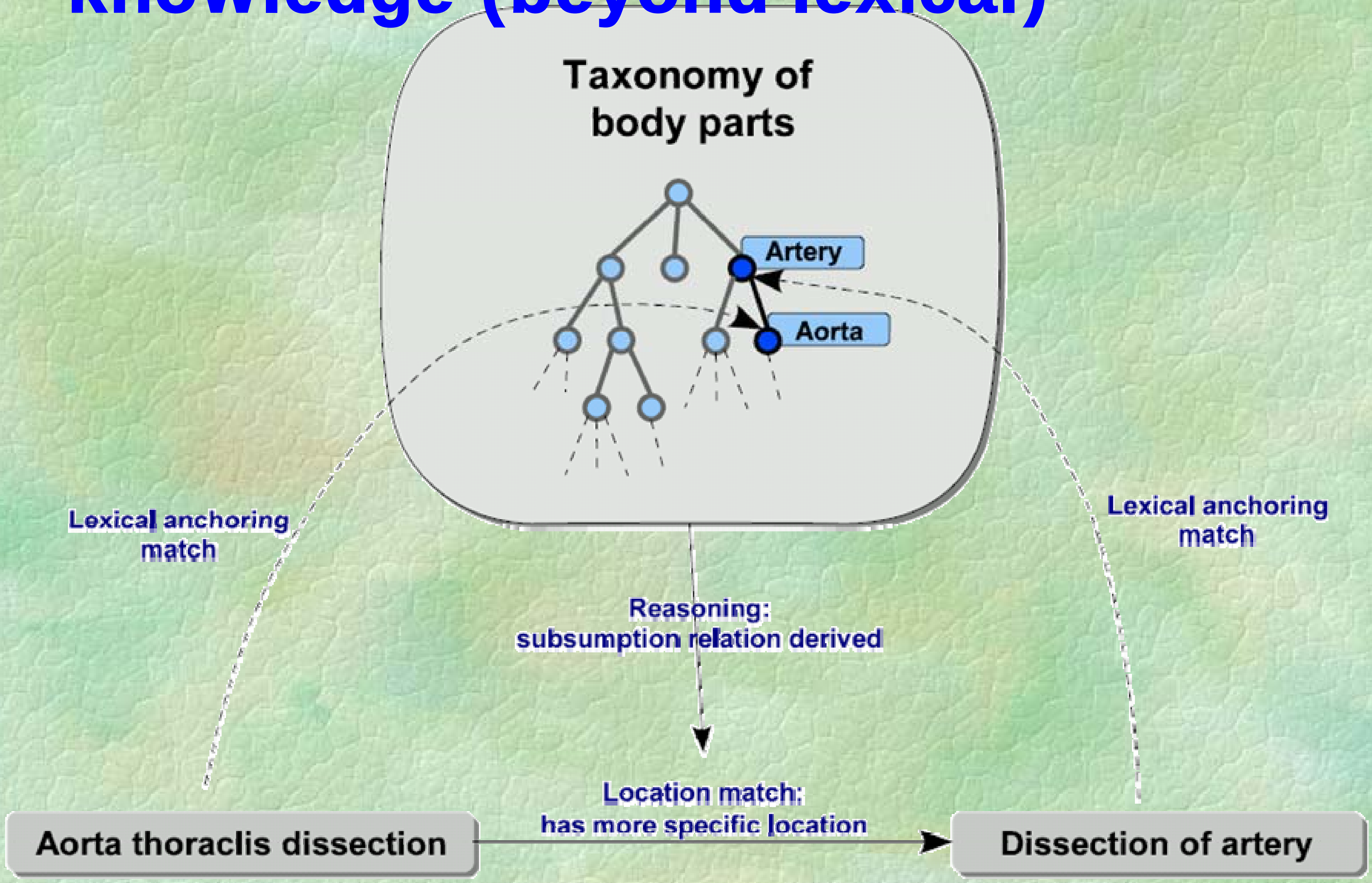
- Combine lexical analysis with hierarchical structure
- First round
  - compare with complete DICE
  - 313 suggested matches, around 70 % correct
- Second round:
  - only compare with “reasons for admission” subtree
  - 209 suggested matches, around 90 % correct

➔ High precision, low recall (“the easy cases”)

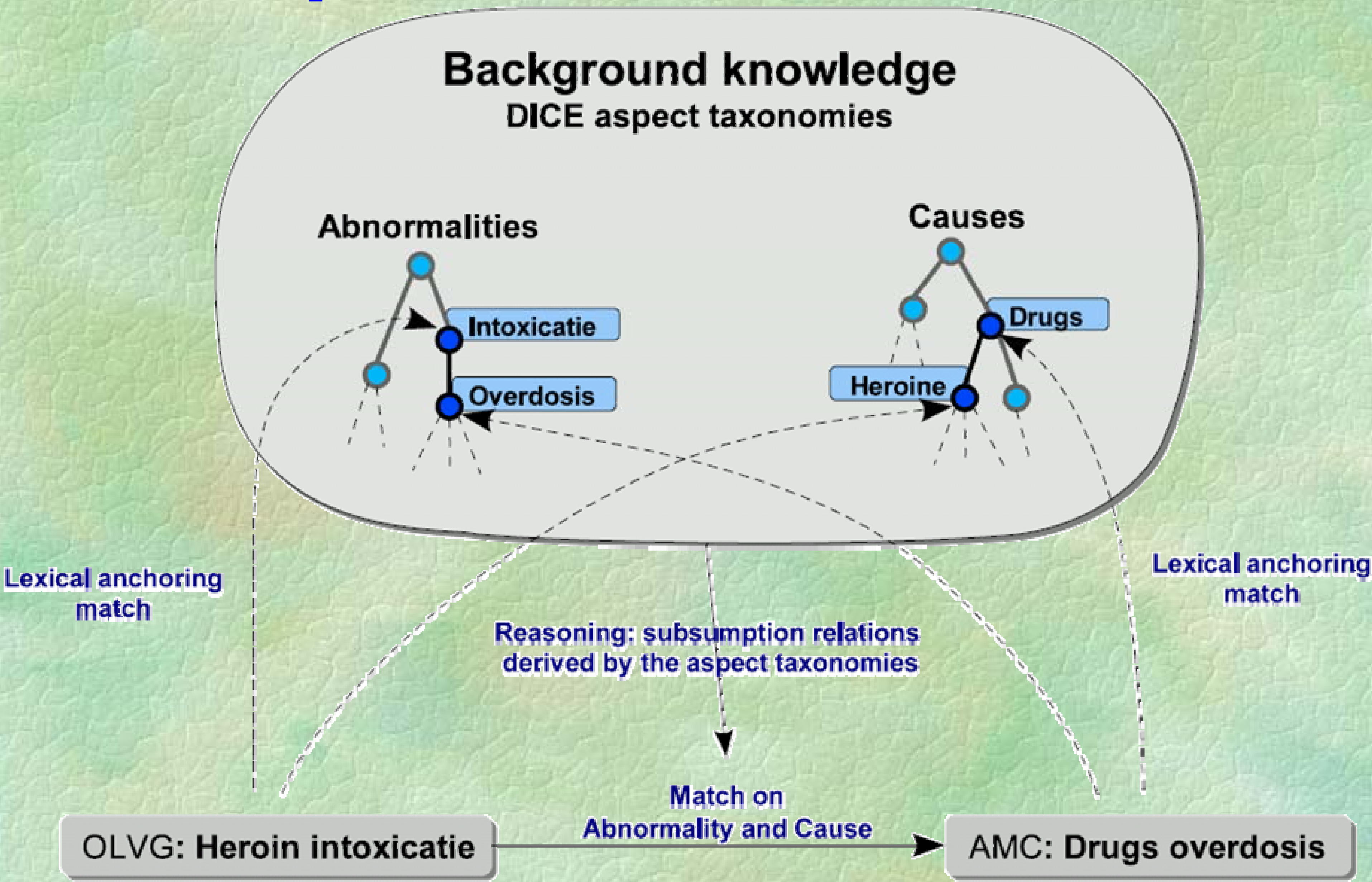
The general idea



Example found with context knowledge (beyond lexical)



Example 2



Background knowledge
DICE aspect taxonomies
Abnormalities
Intoxicatie
Overdosis
Causes
Drugs
Heroine
Lexical anchoring match
Lexical anchoring match
Reasoning: subsumption relations derived by the aspect taxonomies
Match on Abnormality and Cause
OLVG: Heroin intoxicatie
AMC: Drugs overdosis

Anchoring strength

- Anchoring = substring + trivial morphology

anchored on N aspects	OLVG	AMC
N=5	0	2
N=4	0	198
N=3	4	711
N=2	144	285
N=1	401	208
total nr. of anchored terms	549 39%	1404 96%
total nr. of anchorings	1298	5816

Experimental results

- Source & target = **flat lists** of ±1400 ICU terms each
- Background = DICE (2300 concepts in DL)
- Manual Gold Standard (n=200)

	Semantic matching	Own Lexical matching	FOAM	Falcon-AO
agreement on single best match	65 (=32%)	43	35	22
agreement among top 5 matches	8 (= 4%)			
agreement on no match possible	43 (=22%)	43	26	32
improvement over expert match	35 (18%)	6	6	6
TOTAL POSITIVE:	**151 (=76%)**	**92 (=46%)**	**67 (=33%)**	**60 (=30%)**
wrong match found		5	47	78
incorrectly found no match	49(=24%)	103	86	62
TOTAL NEGATIVE:	**49(=24%)**	**108 (=54%)**	**133 (=67%)**	**140 (=70%)**

So…

- The OLVG & AMC terms get their meaning from the context in which they are being used.
- Different background knowledge would have resulted in different mappings
- Their semantics is not context-free
- See also: S-MATCH by Trento

Does more context knowledge help?

Adding more context

1. Only lexical
2. DICE (2300 concepts)
3. MeSH (22000 concepts)
4. ICD-10 (11000 concepts)

- Anchoring strength:

	DICE	MeSH	ICD10
4 aspects	0	8	0
3 aspects	0	89	0
2 aspects	135	201	0
1 aspect	413	694	80
total	548	992	80

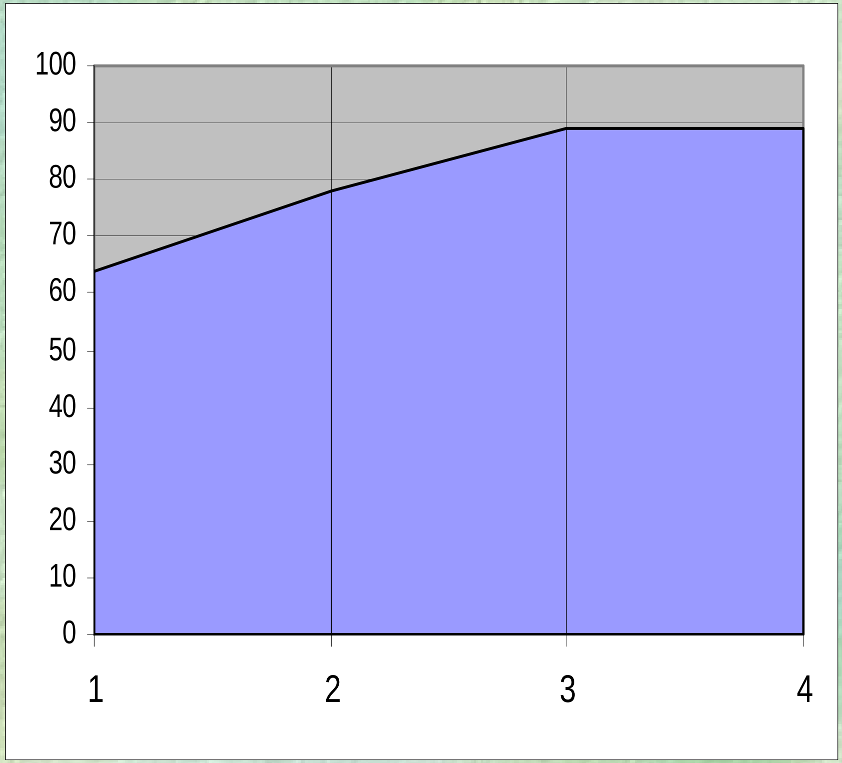
Results with multiple ontologies

Separate

	Lexical	ICD-10	DICE	MeSH
Recall	64%	64%	76%	88%
Precision	95%	95%	94%	89%

Joint

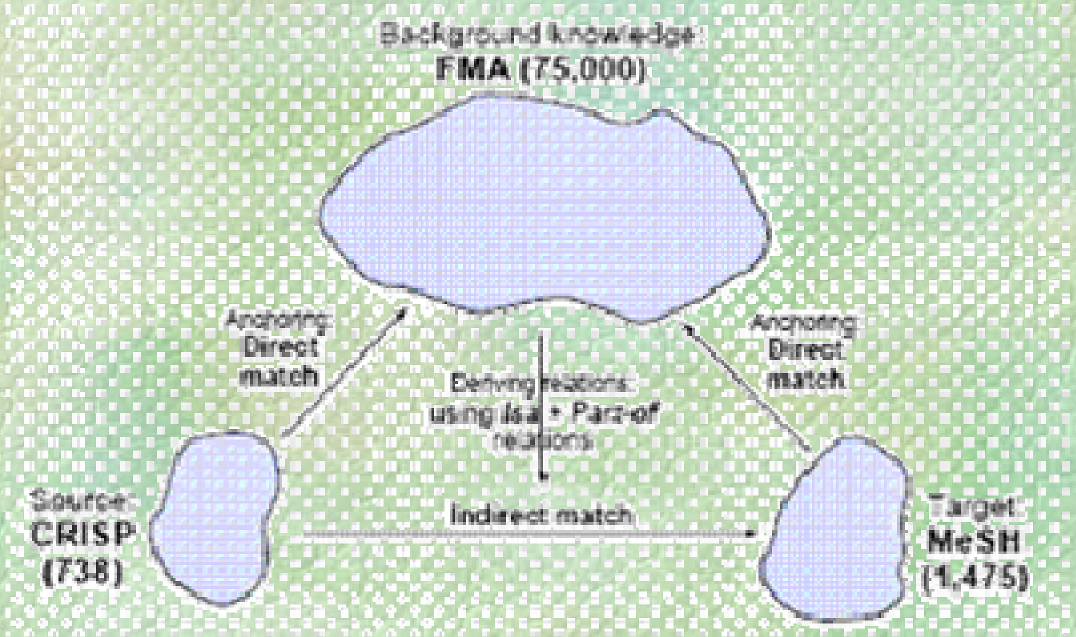
- Monotonic improvement
- Independent of order
- Linear increase of cost



does structured context knowledge help?

Exploiting structure

- CRISP: 700 concepts, **`broader-than`**
- MeSH: 1475 concepts, **`broader-than`**
- FMA: 75.000 concepts, 160 relation-types (we used: **`is-a`** & **`part-of`**)



Direct vs. inferred **matches**

➊ using only:

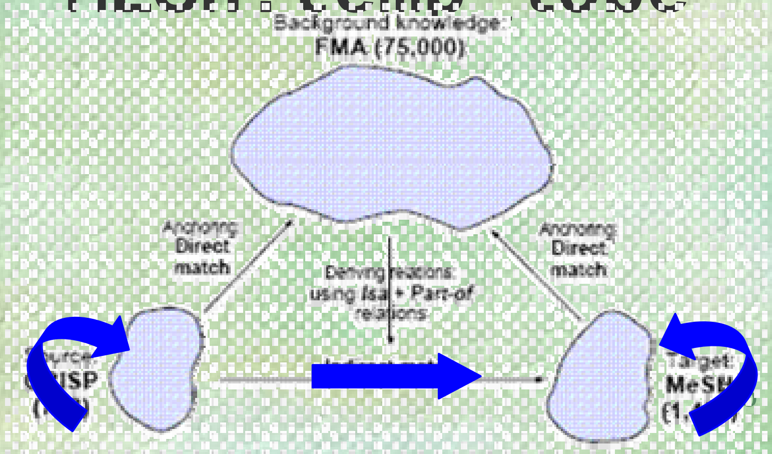
- source-target lexical matches
- relations inside source or target:

e.g: (S $\mathbf{<^d}$ T) & ($\mathbf{T < T'}$) → (S $<^d$ T')

e.g: **`CRISP:brain`** $=^d$ **`MESH:brain`**

`MESH:brain` > **`MESH:temp_lobe`**

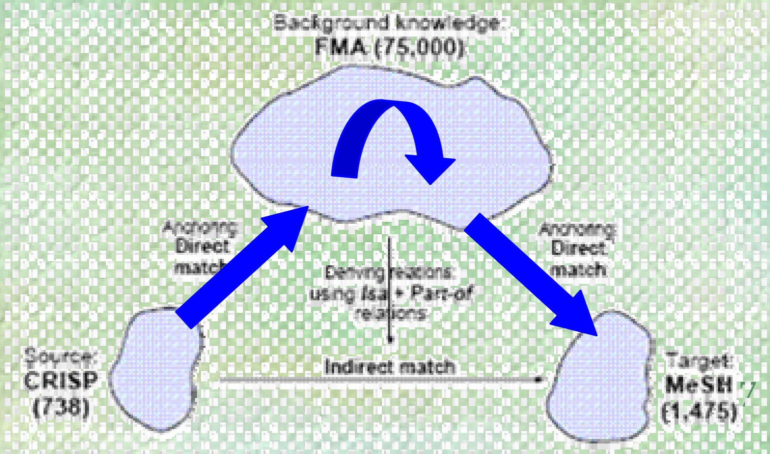
→ **`CRISP:brain`** $>^d$ **`MESH:temp_lobe`**



Direct vs. **inferred matches**

- Using:
  - Lexical anchorings with background
  - Relations inside background knowledge
- Matches inferred via anchorings:

$$(S <^{a} B) \;\&\; (B < B') \;\&\; (B' <^{a} T) \rightarrow (S <^{i} T)$$



Using the structure or not

- $(S <^a B)$ & $(B < B')$ & $(B' <^a T) \rightarrow (S <^i T)$

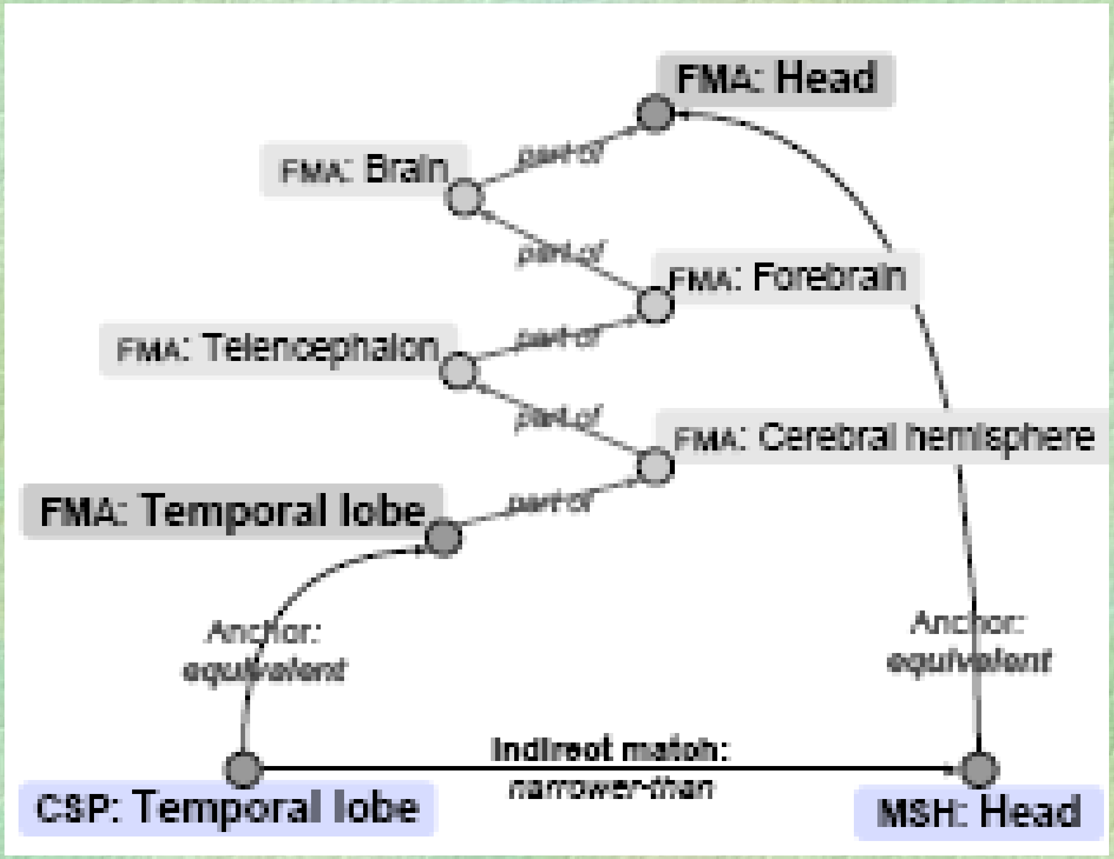
❷ Only stated **`is-a`** & **`part-of`**

❸ Transitive chains of **`is-a`**, and transitive chains of **`part-of`**

❹ Transitive chains of **`is-a`** and **`part-of`**

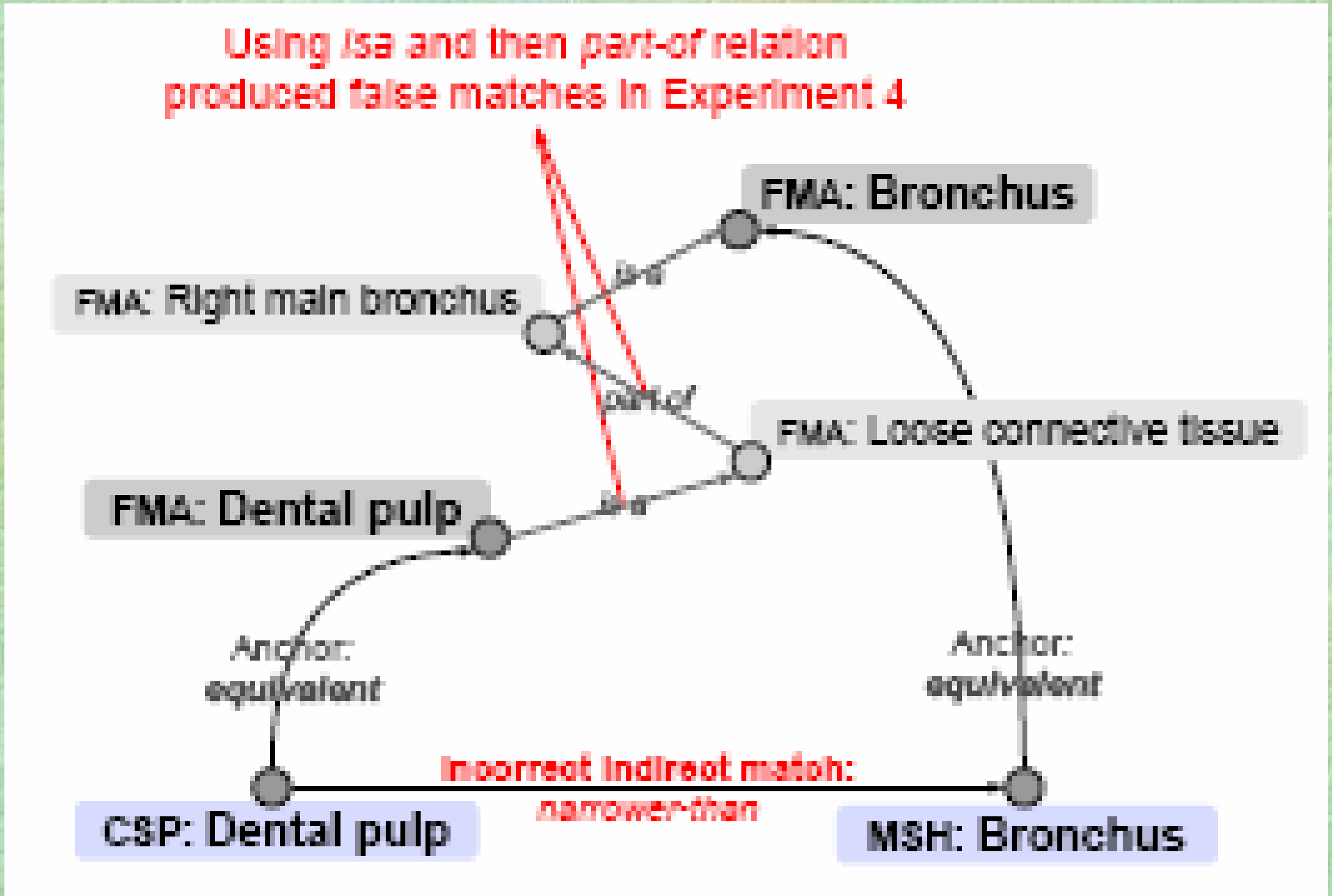
❺ One chain of **`part-of`** before one chain of **`is-a`**

Examples



FMA: Head
FMA: Brain
part-of
part-of
FMA: Forebrain
FMA: Telencephalon
part-of
part-of
FMA: Cerebral hemisphere
FMA: Temporal lobe
part-of
Anchor:
equivalent
Anchor:
equivalent
Indirect match:
narrower-than
CSP: Temporal lobe
MSH: Head

Examples



Using isa and then part-of relation
produced false matches in Experiment 4
FMA: Bronchus
FMA: Right main bronchus
is-a
part-of
FMA: Loose connective tissue
FMA: Dental pulp
is-a
Anchor:
equivalent
Anchor:
equivalent
Incorrect indirect match:
narrower-than
CSP: Dental pulp
MSH: Bronchus

Anchoring strength

	Anchoring concepts	=	≤	≥	Anchored concepts
CRISP to FMA	738	483	607	1474	730
MeSH to FMA	1475	1042	1545	2227	1462

Matching results (CRISP to MeSH)

(Golden Standard n=30)

Recall	=	≤	≥	total	incr.
Exp.1:Direct	448	417	156	1021	-
Exp.2:Indir. is-a + part-of	395	516	405	1316	29%
Exp.3:Indir. separate closures	395	933	1402	2730	167%
Exp.4:Indir. mixed closures	395	1511	2228	4143	306%
Exp.5:Indir. part-of before is-a	395	972	1800	3167	210%

Precision	=	≤	≥	total	correct
Exp.1:Direct	17	18	3	38	100%
Exp.4:Indir. mixed closures	14	39	59	112	94%
Exp.5:Indir. part-of before is-a	14	37	50	101	100%

wrapping up

Related work

- Context knowledge for mapping is mostly linguistic (WordNet)
- Notable exception is S-Match using UMLS, but: we have shown source/target structure is not needed

Conclusions

- Structured investigation on:
  - The role of source/target structure:
    we can even do without, given good context
  - The role of context structure
    (it helps, but be careful with its semantics)
  - The amound of context knowledge
    (surprisingly robust monotonic improvements)