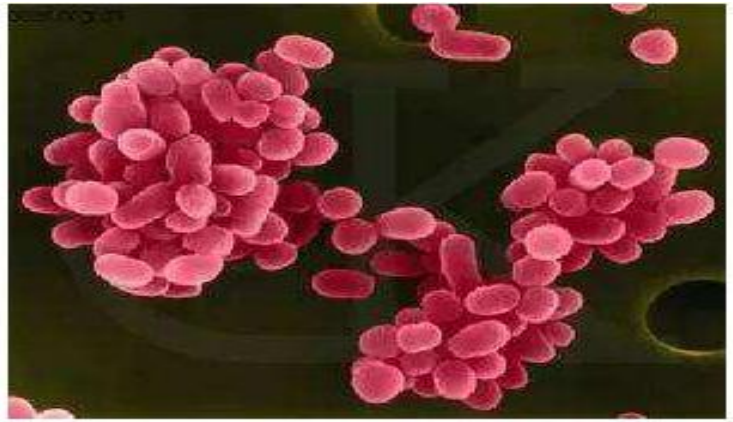
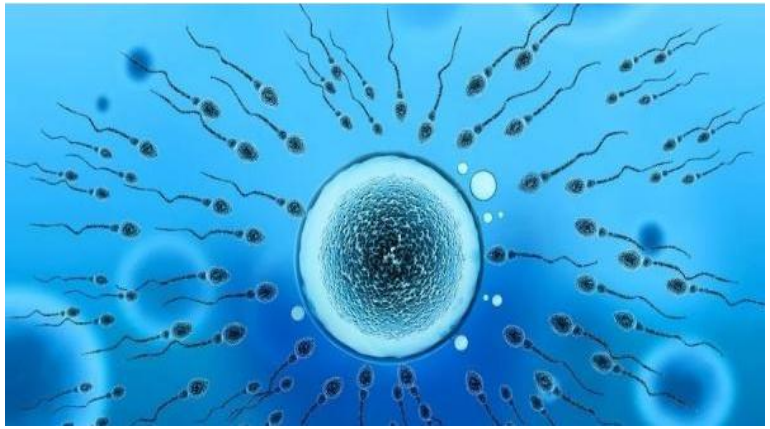
The Curing Terms:
From Medical Terminology to Causal Relations





Xiaowen Wang[12], Natalia Klyueva[1], Emmanuele Chersoni[1], and Chu-Ren Huang[1]

1 Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University

2 School of English and Education, Guangdong University of Foreign Studies

Outline

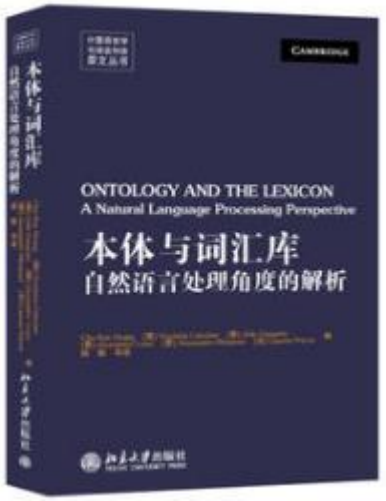
- 1. Introduction
- 2. Literature review
- 3. Methodology
  - 3.1 The corpus
  - 3.2 The initial term list
  - 3.3 Terminology and language ontology
  - 3.4 Terminology and domain ontology
  - 3.5 Extracting causal relations
- 4. Results
- 5. Discussion
- 6. Conclusion
- 7. Future work

Introduction

- Ongoing research combining medical databases and language resources for discovery of causal relations
  - Aiming to learn new information from observational data
  - Focusing on the domain of male infertility

Literature review

From traditional terminology study to ontological terminology



Integration of lexical resources and ontologies

—**Chu-ren Huang**, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, Alessandro Oltramari and Laurent Prévot. ***Ontology and the lexicon: A natural Language Processing Perspective***. Cambridge: Cambridge University Press. 2010.

- **Ontologies** are defined as **"specifications of shared conceptualization"** (Huang, et al., 2010, adapted from Gruber, 1995 and Guarino, 1998).

- Common ontology, domain ontology, linguistic ontology…(Feng, 2005; Huang et al, 2010; Huang and Lee, 2013; Qiu, 2016, etc.)

Medical ontologies

- **Gene Ontology** （GO）
- **Disease and Genotype ontologies**:
- Disease ontology (DO)
- Human Phenotype Ontology (HPO)
- Mammalian Phenotype Ontology (MP)
- Experimental factor ontology (EFO)
- Unified Medical Language System (UMLS)
- SNOMED-CT
- NCI thesaurus
- ICD-10
- LOINC
- …

Causal relations in medical ontologies

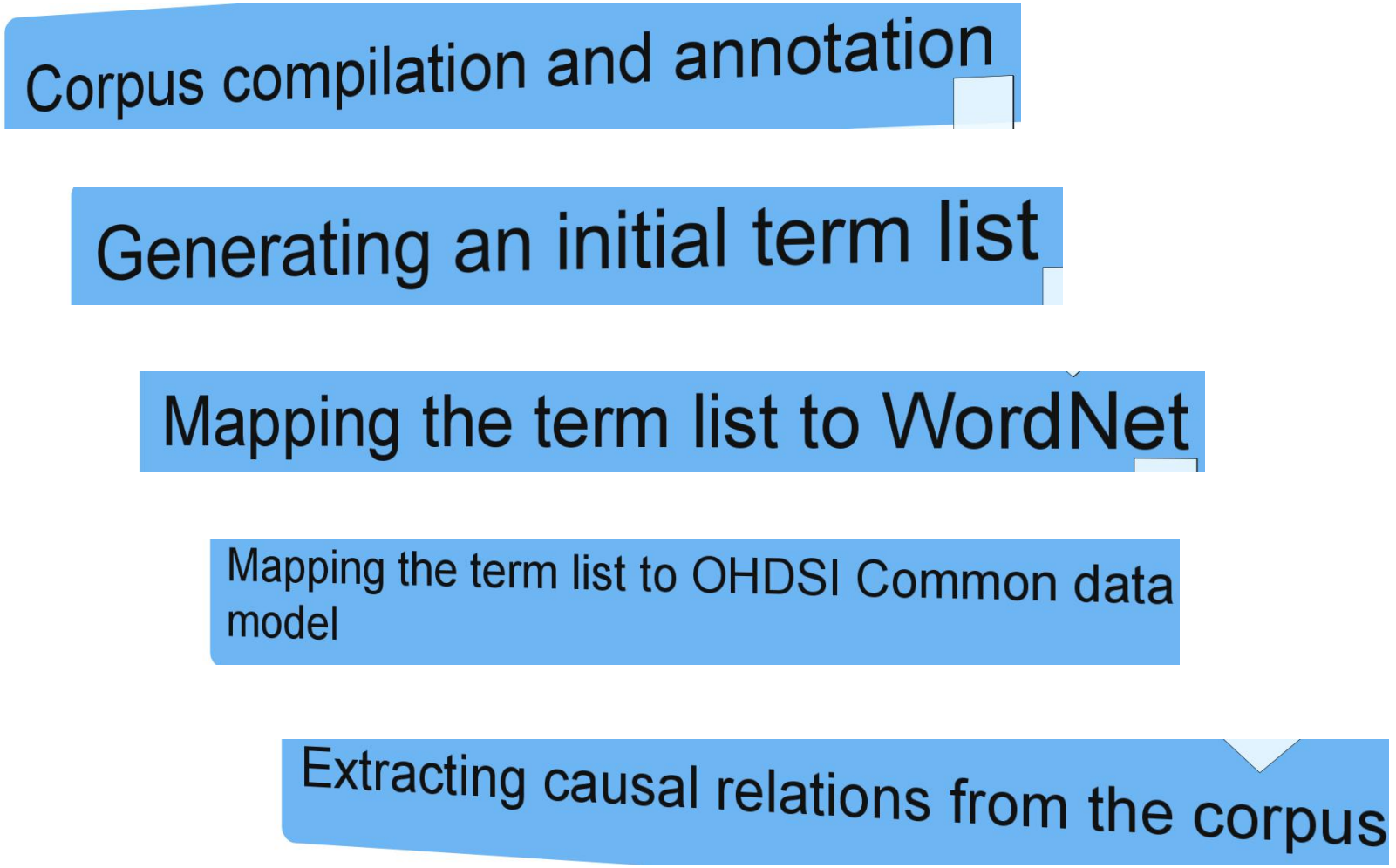
- “What is required is a formalism and ontology capable of dealing with part-whole, **causal** and other transitive relations in medicine plus the relevant content in the concept representation system” (Rector, 1999).
  - AIDS is a viral CNS disorder, unless it has represented within it the fact that AIDS is caused by the HIV virus. This implies significantly more content than is present in most traditional ‘terminologies.’
- “One of the cornerstones of modern medicine is the search for what causes diseases to develop” (Rizzi and Pedersen, 1992)
- ISO 1087 defines **causal relation** as **associative relation "involving cause and its effect".**

Direct causal relations

- Causative agent of
- Bacteria → Carious exposure of pulp

- Four things are required for caries formation: a tooth surface (enamel or dentin), caries-causing bacteria, fermentable carbohydrates (such as sucrose), and time.

Methodology



Corpora

- A self-built corpus of medical research articles in male infertility published in *PLoS ONE* (86 articles, 29385 sentences， 417286 tokens)

- Reference corpora: ***CROWN and CLOB:*** Two Brown Family Corpora Co-built by Chinese Scholars
  - Xu, Jiajin & Maocheng Liang (2012). *Crown: A 2009 Brown family corpus of present-day* ***American English***. National Research Centre for Foreign Language Education, Beijing Foreign Studies University.
  - Xu, Jiajin & Maocheng Liang (2012). *CLOB: A 2009 Brown family corpus of present-day* ***British English***. National Research Centre for Foreign Language Education, Beijing Foreign Studies University.

Table 1. The number of tokens of three generations of Brown family corpora.

	Genre	Sub-corpus tokens	Total tokens		Genre	Sub-corpus tokens	Total tokens
Brown 1961	Fiction	259, 467	1, 027, 021	**LOB 1961**	Fiction	258, 722	1, 018, 785
	General prose	423, 160			General prose	418, 137	
	Learned	163, 309			Learned	162, 322	
	Press	181, 085			Press	179, 604	
Frown 1992	Fiction	260, 414	1, 027, 323	**FLOB 1991**	Fiction	260, 664	1, 024, 643
	General prose	421, 933			General prose	419, 990	
	Learned	163, 228			Learned	163, 286	
	Press	181, 748			Press	180, 703	
Crown 2009	Fiction	259, 250	**1, 026, 226**	**CLOB 2009**	Fiction	259, 484	**1, 023, 466**
	General prose	422, 799			General prose	421, 163	
	Learned	163, 197			Learned	163, 139	
	Press	180, 980			Press	179, 680	

Term list

Generation of the initial term list: keyword list comparing our corpus with Crown+Clob

2036 in total (n.+lexical verbs+adj+adv)

1767 nouns (token)

391 adjectives (token)

387 lexical verbs (token)

48-21(manually filtered) adverbs (token)

Freq	Keyness (LL4)		Effect (DICE)	Keyword	word_PoS_lemma
2599	+	8959.78	0.0124	sperm	sperm_NN_sperm
1443	+	4237.1	0.0069	male	male_JJ_male
1261	+	4419.43	0.006	infertility	infertility_NN_infertility
1171	+	2710.09	0.0056	cells	cells_NNS_cell
1026	+	1576.26	0.0049	men	men_NNS_man
1019	+	2862.36	0.0049	samples	samples_NNS_sample
1018	+	3548.4	0.0049	semen	semen_NN_semen
1010	+	2654.18	0.0048	dna	dna_NN_dna
1000	+	1805.94	0.0048	control	control_NN_control
995	+	2242.9	0.0048	patients	patients_NNS_patient
968	+	2920.89	0.0046	gene	gene_NN_gene
872	+	2254.28	0.0042	expression	expression_NN_expression
819	+	2612.69	0.0039	mice	mice_NNS_mouse
797	+	2311.11	0.0038	genes	genes_NNS_gene
758	+	2687.4	0.0036	infertile	infertile_JJ_infertile
719	+	598.73	0.0034	data	data_NNS_datum
703	+	1910.86	0.0034	protein	protein_NN_protein
694	+	1509	0.0033	cell	cell_NN_cell
626	+	1782.63	0.003	controls	controls_VVZ_control
624	+	946.95	0.003	significant	significant_JJ_significant
622	+	2205.06	0.003	spermatozoa	spermatozoa_NNS_spermatozoon
603	+	2137.68	0.0029	testis	testis_NN_testis
597	+	1302.46	0.0029	normal	normal_JJ_normal
596	+	2076.2	0.0029	pcr	pcr_NN_pcr
576	+	931.74	0.0028	levels	levels_NNS_level
573	+	1821.11	0.0027	males	males_NNS_male
541	+	1023.93	0.0026	associated	associated_VVN_associate
516	+	1774.61	0.0025	testes	testes_NNS_testis

Freq	Keyness (LL4)		Effect (DICE)	Keyword	word_PoS_lemma
2599	+	8959.78	0.0124	sperm	sperm_NN_sperm
1261	+	4419.43	0.006	infertility	infertility_NN_infertility
1171	+	2710.09	0.0056	cells	cells_NNS_cell
1026	+	1576.26	0.0049	men	men_NNS_man
1019	+	2862.36	0.0049	samples	samples_NNS_sample
1018	+	3548.4	0.0049	semen	semen_NN_semen
1010	+	2654.18	0.0048	dna	dna_NN_dna
1000	+	1805.94	0.0048	control	control_NN_control
995	+	2242.9	0.0048	patients	patients_NNS_patient
968	+	2920.89	0.0046	gene	gene_NN_gene
872	+	2254.28	0.0042	expression	expression_NN_expression
819	+	2612.69	0.0039	mice	mice_NNS_mouse
797	+	2311.11	0.0038	genes	genes_NNS_gene
719	+	598.73	0.0034	data	data_NNS_datum
703	+	1910.86	0.0034	protein	protein_NN_protein
694	+	1509	0.0033	cell	cell_NN_cell
622	+	2205.06	0.003	spermatozoa	spermatozoa_NNS_spermatozoon
603	+	2137.68	0.0029	testis	testis_NN_testis
596	+	2076.2	0.0029	pcr	pcr_NN_pcr
576	+	931.74	0.0028	levels	levels_NNS_level
573	+	1821.11	0.0027	males	males_NNS_male
524	+	1857.53	0.0025	daz	daz_NP_daz
516	+	1774.61	0.0025	testes	testes_NNS_testis
489	+	714.21	0.0023	cases	cases_NNS_case
489	+	814.48	0.0023	type	type_NN_type
466	+	1616.7	0.0022	methylation	methylation_NN_methylation
452	+	1577.29	0.0022	deletion	deletion_NN_deletion
450	+	724.28	0.0022	test	test_NN_test
422	+	1482.15	0.002	spermatogenesis	spermatogenesis_NN_spermatogenesis

The language ontology: Mapping with WordNet

- Cleaning → lemmatization → mapping
- Examples:
- ========== sperm ==========
- **Definition**: the male reproductive cell; the male gamete
- **Examples**: ['a sperm is mostly a nucleus surrounded by little other cellular material']
- **Synonyms**: {'spermatozoan', 'sperm', 'spermatozoon', 'sperm_cell'}
- **Hyponyms**: []
- **Hypernyms**: [Synset('gamete.n.01'), Synset('reproductive_cell.n.01'), Synset('cell.n.02'), Synset('living_thing.n.01'), Synset('whole.n.02'), Synset('object.n.01'), Synset('physical_entity.n.01'), Synset('entity.n.01')]
- ========== infertility ==========
- **Definition:** the state of being unable to produce offspring; in a woman it is an inability to conceive; in a man it is an inability to impregnate
- **Examples**: []
- **Synonyms**: {'infertility', 'sterility'}
- **Hyponyms**: [Synset('barrenness.n.01'), Synset('cacogenesis.n.01'), Synset('dysgenesis.n.01'), Synset('impotence.n.02'), Synset('erectile_dysfunction.n.01')] Hypernyms: [Synset('physical_condition.n.01'), Synset('condition.n.01'), Synset('state.n.02'), Synset('attribute.n.02'), Synset('abstraction.n.06'), Synset('entity.n.01')]

The domain ontology: Mapping with OHDSI Common Data Model (CDM)

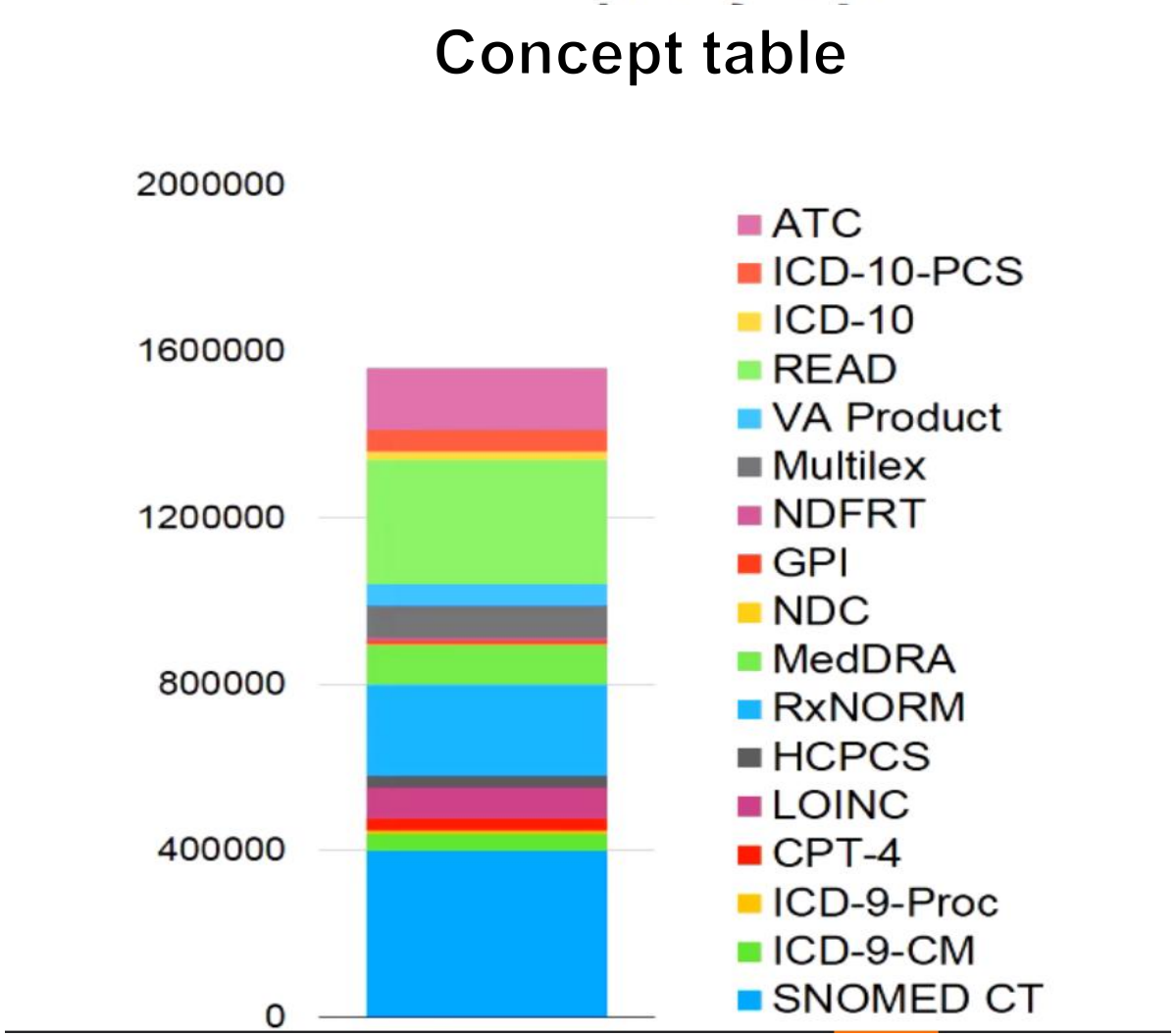
- **Common Data Model**

What it is

- Standardized structure to house existing vocabularies used in the public domain
- Compiled standards from disparate public and private sources and some OMOP-grown concepts

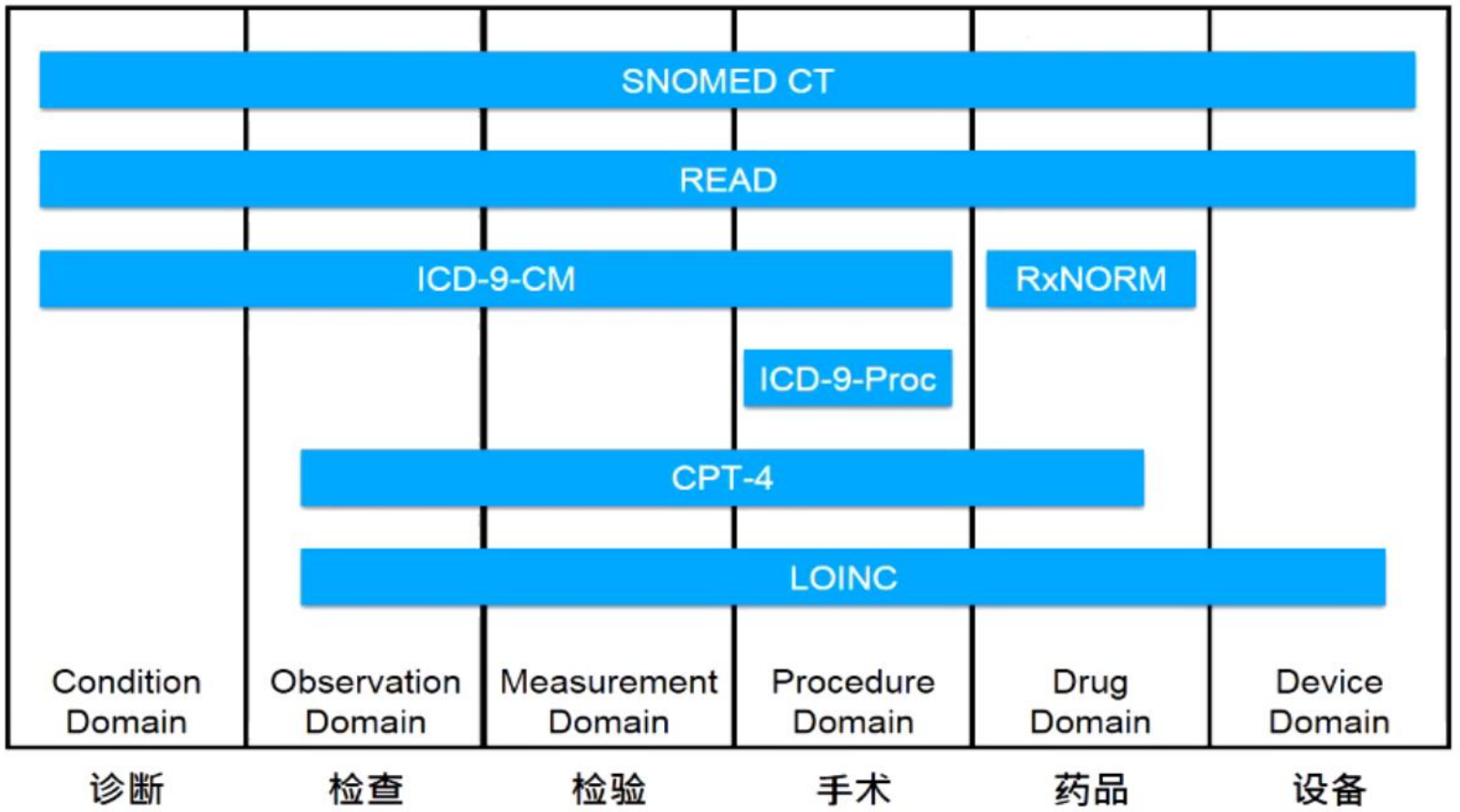
What it's not

- Static dataset – the vocabulary updates regularly to keep up with the continual evolution of the sources
- Finished product – vocabulary maintenance and improvement is ongoing activity that requires community participation and support



Concepts in the Common Data Model are derived from a number of public or proprietary terminologies such as SNOMED-CT and RxNorm, or custom generated to standardize aspects of observational data.

Domain和Vocabulary



Extracting causal relations from the corpus

- Method 1: A PMI-based algorithm

A simple, PMI-based algorithm for extracting relations on the basis of the "cause" keyword from the male infertility corpus

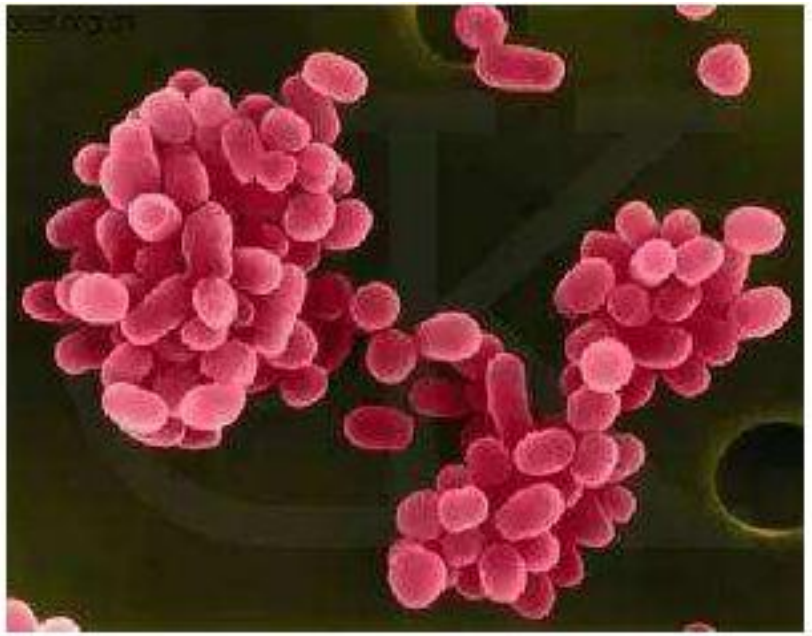
- 1) it stores the frequencies for the nouns of the initial term list, as well as of all the other nouns occurring in the corpus;
- 2) it stores the frequencies of the co-occurrence of the lemmatized word “cause” with every single noun in the corpus, and records the triples in the form NOUN1 cause NOUN2;
- 3) it computes the mutual information between "cause" and all the preceding and the following nouns. For the triples included in our results, the only constraint is that at least one of the nouns has to be included in the initial term list.
- output：427 related pairs

➢Other constructions indicating causal relationship which we call “core pattern”: induced by, etiology of, due to, etc.

Haemophilus **parainfluenzae** is as an opportunistic pathogen which causes systemic diseases including endocarditis, meningitis and bacteremia, and is often isolated from the sputa of patients with chronic obstructive lung disease.

haemophilus-n	cause	meningitis-n	7.753402799
haemophilus-n	cause	sputum-n	7.753402799
haemophilus-n	cause	endocarditis-n	7.753402799
haemophilus-n	cause	bacteremia-n	7.060255618

- Haemophilus could potentially be markers for future clinical applications and investigations of male infertility.



Method 2: relation extraction from the corpus based on syntactic patterns

annotation

acl : clausal modifier of noun (adjectival clause)
advcl : adverbial clause modifier
advmod : adverbial modifier
amod : adjectival modifier
appos : appositional modifier
aux : auxiliary
case : case marking
cc : coordinating conjunction
ccomp : clausal complement
clf : classifier
compound : compound
conj : conjunct
cop : copula
csubj : clausal subject
dep : unspecified dependency
det : determiner

discourse : discourse element
dislocated : dislocated elements
expl : expletive
fixed : fixed multiword expression
flat : flat multiword expression
goeswith : goes with
iobj : indirect object
list : list
mark : marker
nmod : nominal modifier
nsubj : nominal subject
nummod : numeric modifier
obj : object
obl : oblique nominal
orphan : orphan
parataxis : parataxis
punct : punctuation
reparandum : overridden disfluency
root : root
vocative : vocative
xcomp : open clausal complement

noun_subject		noun_object
extra-germination	cause	infertility
use	cause	infertility
contraceptives	caused	infertility
overload	causes	infertility
mutations	cause	infertility
beverages	cause	infertility
mutilation	cause	infertility
operations	cause	infertility

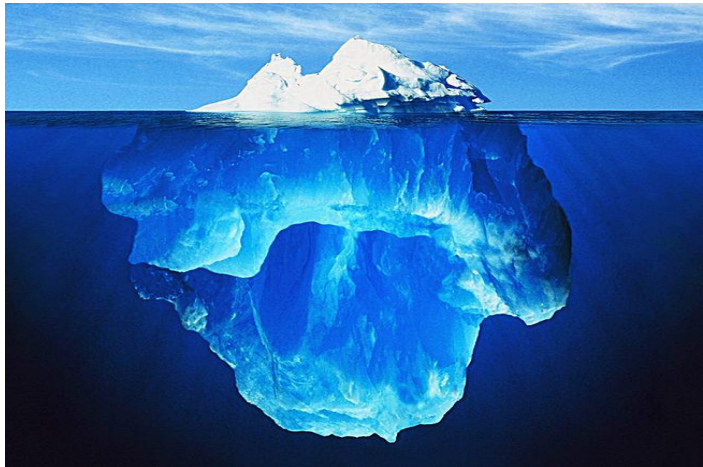
Discussion:
results of relation extraction from the corpus

noun_subject		noun_object	concordance line
extra-germination	cause	infertility	At the opposite side of this is the belief that an extra-germination of the clitoris could cause infertility.
use	Cause	Infertility	The perception that the use of contraceptives can cause infertility in females therefore poses a challenge to approach adopted by the Ghana Health Service in her reproductive health policy to prevent unwanted pregnancies and ensure good child and maternal
overload	Causes	Infertility	This work shows that an overload of dietary cholesterol causes complete infertility in dyslipidemic male mice (the Liver X Receptor-deficient mouse model).

mutations	Cause Infertility	The loss-of-function mutations of these genes might not only cause infertility but also testicular tumors and other related diseases [21] .
mutilation	Cause Infertility	The belief that Female Genital Mutilation (FGM) can cause infertility in females is of medical importance.
operations	Cause Infertility	Female pelvic infection can cause pelvic inflammatory disease and thus the occurrence of pelvic adhesions, resulting in infertility [25], and negative-pressure operations during abortions may also cause immune infertility.

Future work

- To extend the scope of study:
  - from nouns to verbs, adjectives, adverbs, etc.
  - from single words to lexical bundles
- To enrich information of medical causal relations with qualia eventive knowledge
- Error analysis of the relation extraction, comparing strength and weakness of each method
- To combine the methods to achieve the best optimal way of enriching causal information
- Deep learning
- **Word-embedding**



Word-embedding

```
model = Doc2Vec.load("plosone_doc2vec.vec")
```

```
model.most_similar(positive=['cancer', 'breast'], negative=['man'])
```

```
/home/qt402/.local/lib/python3.5/site-packages/ipykernel_launcher.py:1: DeprecationWarning: Call to deprecated `most_simi
lar` (Method will be removed in 4.0.0, use self.wv.most_similar() instead).
  """Entry point for launching an IPython kernel.
```

```
[('prostate', 0.6743712425231934),
 ('cancer.', 0.6500186324119568),
 ('cancer,', 0.620056688785553),
 ('quality-specific', 0.61820387840271),
 ('endometrial', 0.6146716475486755),
 ('118).', 0.6074773073196411),
 ('cancers', 0.5930484533309937),
 ('colorectal', 0.5908399820327759),
 ('Jade', 0.5904192328453064),
 ('melanoma', 0.5845313668251038)]
```

Conclusion

- The ongoing study combines medical databases and language resources for discovery of causal relations
  - mapping the term list to linguistic and domain ontological resources
  - extending it by extracting causal relations from self-built corpus
- Analysis of data from our pilot study confirms that our approach could lead to potential discovery of new causal relations.
- We also found that causal relations reported in observational data as well as medical papers are often descriptive causal relations and not logical direct causal relations.
- Such information can be enriched by combination with qualia eventive knowledge.

References

- Chu-ren Huang, Nicoletta Calzolari, Aldo Gangemi, Alessandro Lenci, Alessandro Oltramari and Laurent Prévot. (2010). *Ontology and the lexicon: A natural Language Processing Perspective*. Cambridge: Cambridge University Press.
- Chu-Ren Huang, and Sophia Y.M. Lee. 2013. 知识的系统与知识系统的建构：绪论知识本体语言科学整合研究. In Chu-Ren Huang and Sophia Y. M. Lee, Eds. Special Issues on Ontology and Chinese Language Processing. Contemporary Linguistics (当代语言学). 2013.3
- Chunyu Kit, Zhiwei Feng. (2009). Ontology-Based Definition of Term (2). *Terminology Standardization & Information Technology (术语标准化与信息技术)*. pp. 4-8， 43.
- Chunyu Kit, Zhiwei Feng. (2009). Ontology-Based Definition of Term (3). *Terminology Standardization & Information Technology (术语标准化与信息技术)*. pp. 14-23.
- DA Rizzi , SA Pedersen . (1992). Causality in medicine: towards a theory and terminology. *Theor Med*. Sep;13(3):233-54.
- Eugen Wüster. (1991). Einführung in die allgemeine Terminologie und terminologische Lexikographie，3． Auflage，Bonn：Romanistischer Verlag， .
- International Standard ISO 704 < Terminology work - Principles and methods > (2009)．https://www.iso.org/standard/38109.html
- Zhiwei Feng. (2005). On Humanity Spirit of Natural Language Processing from the Viewpoint of Ontology（从知识本体谈自然语言处理的人文性）. Applied Linguistics（语言文字应用）, (4).

Thank you !

Xiaowen Wang[12], Natalia Klyueva[1], Emmanuele Chersoni[1], and Chu-Ren Huang[1]

1 Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University

2 School of English and Education, Guangdong University of Foreign Studies

xiaowen-annie.wang@connect.polyu.hk

natalia.klyueva@polyu.edu.hk

emmanuelechersoni@gmail.com

churen.huang@polyu.edu.hk