Komenti: semantic query and text mining framework

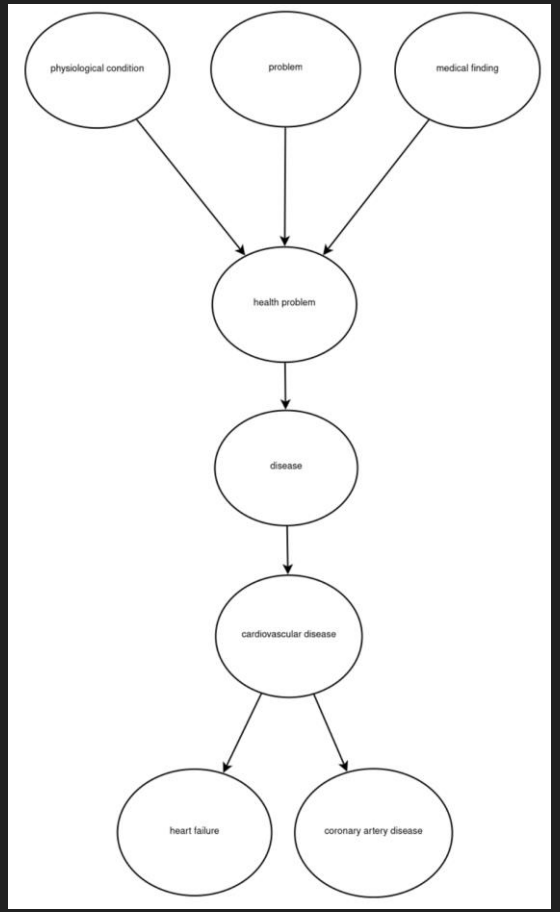
Luke Slater (l.slater.1@bham.ac.uk)
HDR UK National Text Analytics workshop
12/03/2021

Objectives

- Background and motivation
- Design
- Practical examples
- On-going & Future

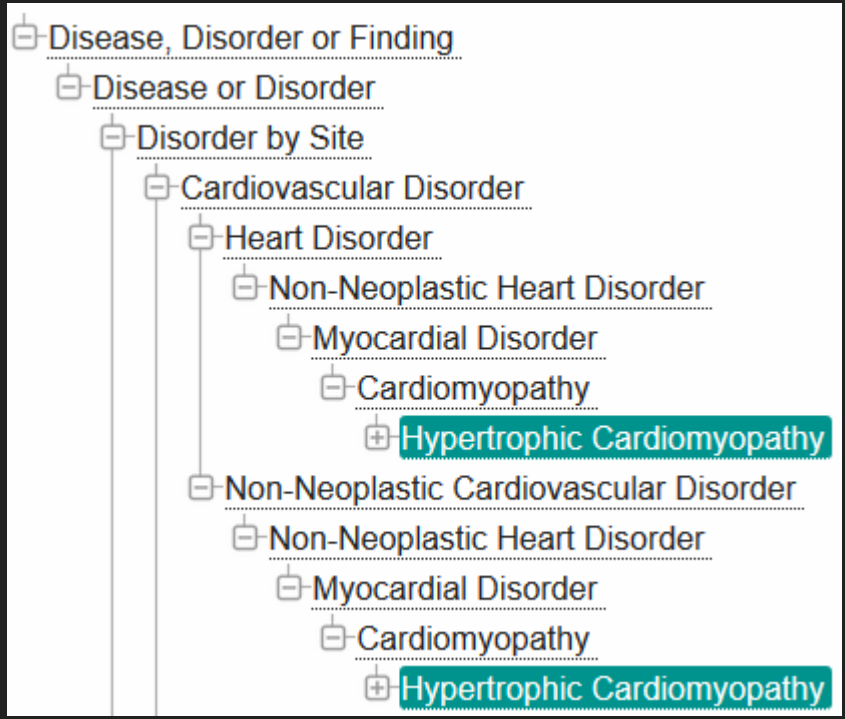
Ontologies: Describing Kinds of Objects

- Ontologies categorise, define and relate the kinds of things in a domain.
- Easily thought of as a 'schema' for objects
- Exact definitions differ, but most ontologies share four main features:
  - Classes and relations.
  - Domain vocabulary.
  - Metadata and descriptions.
  - Axioms and formal definitions.



Hierarchies and Controlled Domain Vocabularies

- Controlled domain vocabularies are authoritative resources for named entities in biomedical fields, containing metadata including synonyms
- They are usually found in ontologies, which also have other features, such as hierarchy and formal logic axioms
- Establishing computational consensus
- Query groups of terms



Motivations

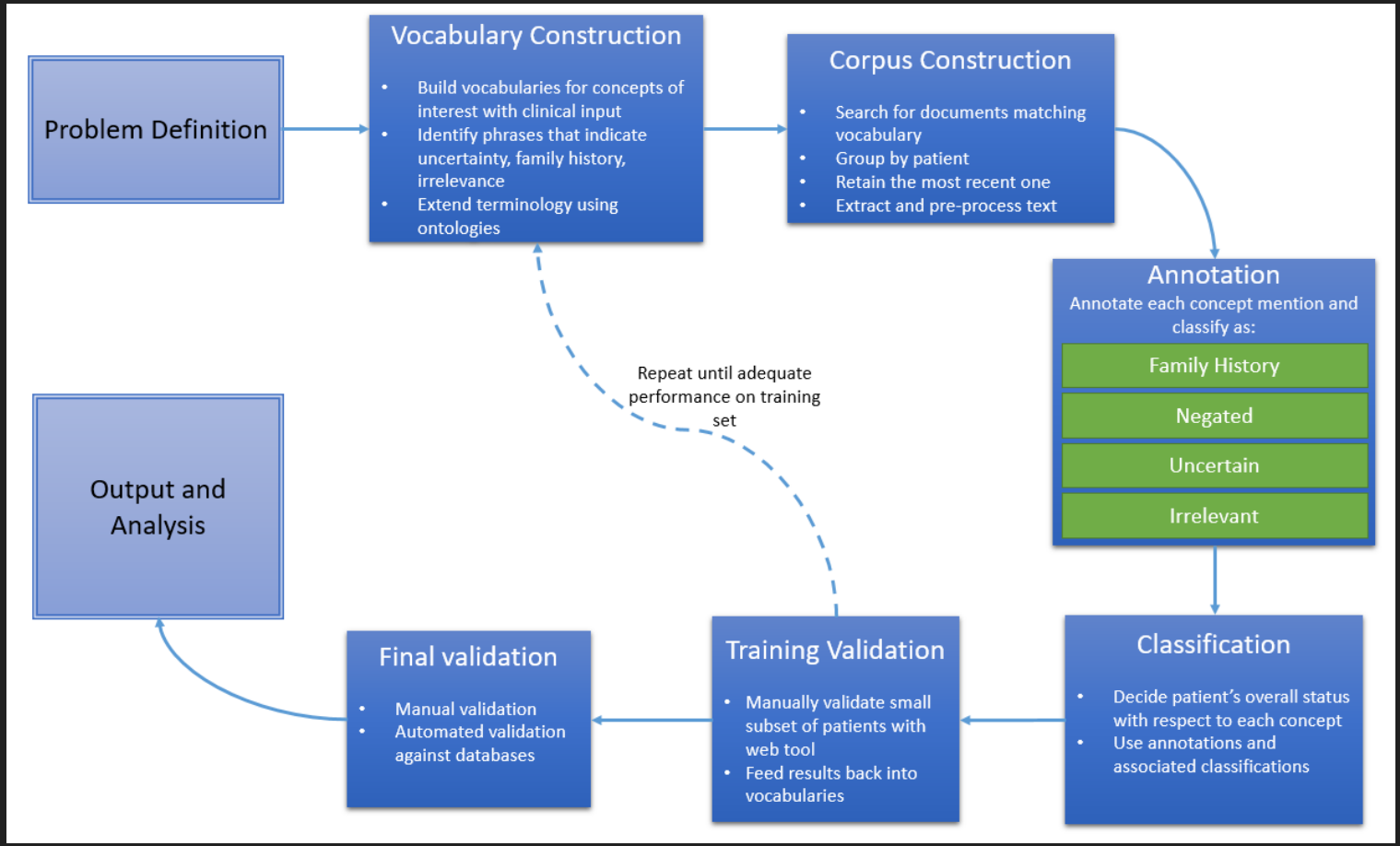
- The problems ontologies provide solutions for map to the problems (sources of error) for text mining:
  - Variability: Human language describes things in different ways (e.g. HCM vs HOCM vs hypertrophic cardiomyopathy vs unexplained thickening of the heart)
  - Ambiguity: different mentions of the same thing may be more or less specific (e.g. cardiovascular disease vs heart failure vs diastolic cardiac insufficiency)[1]
- How can we make full use of these features? Can we use the /other/ features of ontologies also???

Komenti: Features

- Semantic query and text mining framework
- Novel methods:
  - Method of drawing vocabulary from all 350+ biomedical ontologies, instead of only one[1]
  - Ability to query groups of terms with description logic queries[2]
  - Novel negation, uncertainty detection algorithm[3]
  - Automated ontology extension through text mining for improved classification performance4
  - Experimental automated description logic axiom determination from text...
  - Experimental semantic similarity features…[5]

[1] Slater LT, Bradlow W, Ball S, Hoehndorf R, Gkoutos GV. Improved characterisation of clinical text through ontology-based vocabulary expansion. Journal of Biomedical Semantics. 2021. Accepted: In Press.
[2] Slater LT, Bradlow W, Hoehndorf R, Motti DF, Ball S, Gkoutos GV. Komenti: A semantic text mining framework. bioRxiv. 2020 Aug 4;2020.08.04.233049.
[3] Slater LT, Bradlow W, Motti DFA, Hoehndorf R, Ball S, Gkoutos GV. A fast, accurate, and generalisable heuristic-based negation detection algorithm for clinical text. Computers in Biology and Medicine. 2021 Mar 1;130:104216.
[4] Slater LT, Hoehndorf R, Karwath A, Gkoutos GV. Exploring Binary Relations for Ontology Extension and Improved Adaptation to Clinical Text. bioRxiv. 2020 Dec 4;2020.12.04.411751.
[5] Slater LT, Karwath A, Williams JA, Russell S, Makepeace S, Carberry A, et al. Towards Similarity-based Differential Diagnostics For Common Diseases. bioRxiv. 2021 Jan 27;2021.01.26.428269.

Komenti



[1] Slater LT, Bradlow W, Hoehndorf R, Motti DF, Ball S, Gkoutos GV. Komenti: A semantic text mining framework. bioRxiv. 2020 Aug 4; 2020.08.04.233049.

Use Example 1: HCM Cohort Discovery

- The hospital has a rare disease registry for hypertrophic cardiomyopathy (HCM) patients
- We suspect there are many HCM patients known to the hospital, unknown to structured resources
- Process:
  - Build a vocabulary for hypertrophic cardiomyopathy
  - Annotation of letters and discernment of mention context
  - ‘Diagnose’ those patients HCM using that evidence, classification step
  - Expert validation
- When we did this, we found 861 patients with HCM unknown to the specialist registry or ICD codes, and a further 696 patients without HCM who had a family history of it[1]

[1] Slater LT*, Bradlow W*, Desai T, Aziz A, Evison F, Ball S, Gkoutos GV. Computerised Identification of Patients Using Routine Clinical Records; Towards Population Health Management in Hypertrophic Cardiomyopathy. Under Preparation.

Use Example 2: Medication Audit

- We know that patients with HCM should be anti-coagulated if they have atrial fibrillation, and we want to find out if this is the case for all known patients…
- Process:
  a. Build a vocabulary for atrial fibrillation and anti-coagulant drug names (a list of phrases to match in text)
  b. Annotate patients, and associate ontology terms with each
  c. ‘Diagnose’ those patients for atrial fibrillation and anti-coagulant drug status
  d. Expert validation
  e. From this we can find patients who have atrial fibrillation, but aren’t being anti-coagulated
- When we actually did this, we did find 6 patients who were not anti-coagulated. They were referred to the specialist for improved treatment...[1]
- In this way we can start to test our clinical hypotheses on text records...

[1] Slater LT, Bradlow W, Hoehndorf R, Motti DF, Ball S, Gkoutos GV. Komenti: A semantic text mining framework. bioRxiv. 2020 Aug 4; 2020.08.04.233049.

Use Example 3

- Instead of simply picking up already known diagnoses from text...
- Automated diagnosis across many diseases from text-derived phenotypes
- Uses in:
  - Automated coding
  - Differential diagnosis
  - Outcome prediction
  - Document classification
  - Other stratification

Phenotype Profile

- Phenotype profiles are a list of phenotype classes associated with an entity
- Definitional phenotype profiles vs patient phenotype profiles
  - Sources such as MIMIC, or derived from co-ocurrence matrices

Herniated disc:

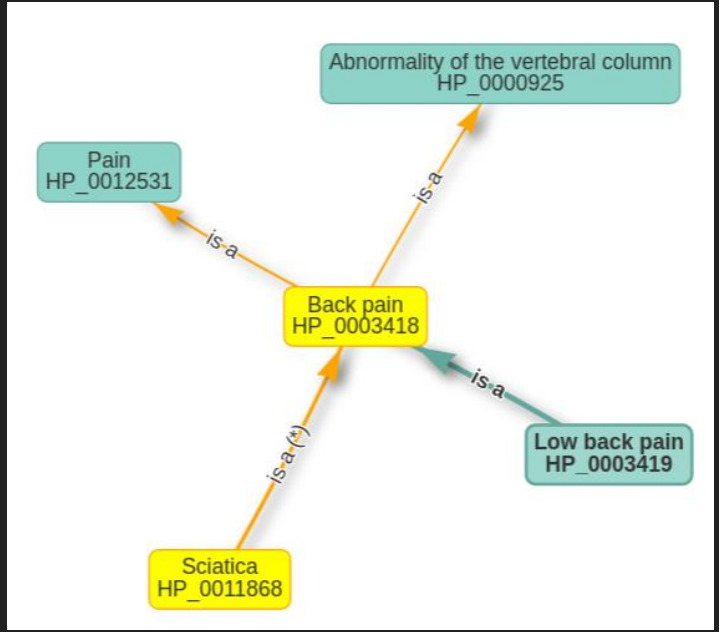
- Sciatica
- Abnormal gait
- Difficulty sleeping

Patient X:

- Lower back pain
- Difficulty sleeping
- Insomnia
- Difficulty walking
- Anxiety

Semantic Similarity

- Uses the taxonomy as a graph structure to calculate a measure of similarity between two phenotype profiles (e.g. a patient phenotype and a disease phenotype)
- Individual phenotypes are compared based on how semantically close they are per the ontology
- This allows you to make similarity-based rankings of biomedical entities (e.g. patients to disease profiles, or patients to other patients)

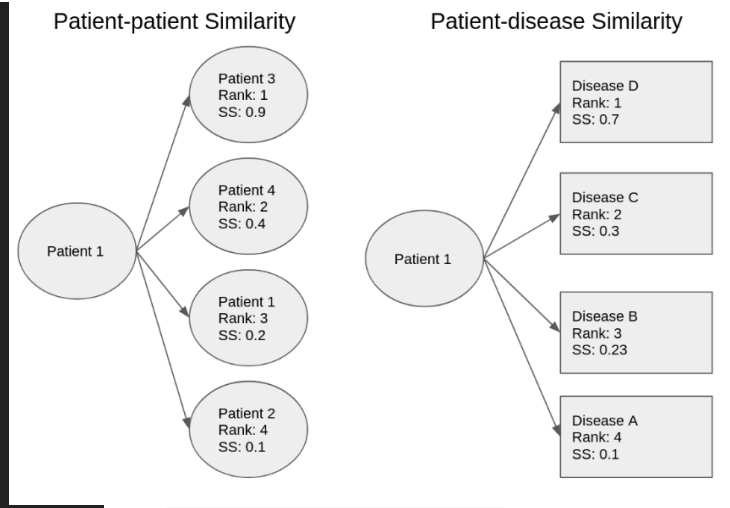


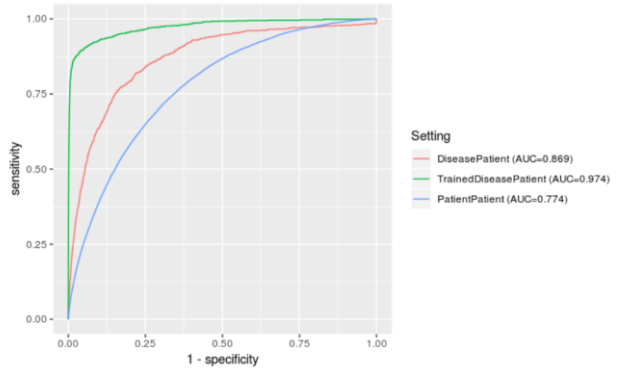
Use Example 3

- Sampled 1,000 patient visits from MIMIC-III
- Annotated the texts with HPO, producing a phenotype profile for each patient visit
- Tested three methods of prioritising their primary disease:
  - Comparing each patient profile with the other text-derived patient profiles
  - Comparing each patient profile with text-mined literature abstract derived HP phenotypes for diseases in DO
  - Extended literature-derived phenotypes using text-mined patient phenotype profiles mined from a training set
- In both modes you get a ranked list of similar entities for each patient visit
- Evaluated how well this was predictive of their primary diagnosis

Table 1. Performance for matching first diagnosis of MIMIC patients under different settings. Top ten accuracy is the percentage of patients for whom a correct diagnosis appeared in the ten most similar entities.

Setting	AUC	MRR	Top Ten Accuracy
Patient Comparison	0.774 (0.7724-0.7762)	0.423	0.606
Disease Comparison	0.869 (0.8564-0.8818)	0.016	0.029
Trained Disease Comparison	0.974 (0.9682-0.9796)	0.314	0.638





[1] Slater LT, Karwath A, Williams JA, Russell S, Makepeace S, Carberry A, et al. Towards Similarity-based Differential Diagnostics For Common Diseases. bioRxiv. 2021 Jan 27;2021.01.26.428269.

Current use cases

- Moving towards similarity-based classification in a direct clinical setting at our hospital...
- Identification and characterisation of patients with rare diseases from text records
- Prediction of death in critical care using initial text record
- Prediction of outcomes in rare diseases
- Automated medicines audit and risk scoring
- Web-based tool for patient scheduling using integration of structured and text-mined data (in collaboration with our hospital’s IT dept) for clinician-directed cohort identification

Conclusions

- Komenti aims to make full use of the semantic features of biomedical ontologies
- It has a large set of features to this effect
- It has led to improved clinical outcomes, with experiments and implementation continuing
- Thank you, questions welcome