Machine Learning guided and manual curation of an Ontology for Uveitis

Samantha C Pendleton

Centre of Computational Biology, University of Birmingham

BACKGROUND

- Ontologies represent a domain of knowledge with all its logic via classes, relationships, and annotations (e.g. synonyms & cross-references to other ontologies) [1].
- In the 1970s, AI researchers argued that they could create new ontologies via AI, however by the 1980s they were prevented by AI Winter. We are now in an era that AI in thriving and so we come back to this initial hypothesis.
- To begin our automated ontology path, we must start with semi-automated building with medical and public data: to combine the mixture of medical & natural language (synonyms).
- Uveitis is a collection of phenotype characterised by eye inflammation, there are different types of Uveitis and for each there are different treatments, complications [2].
- The available biomedical ontologies for Uveitis are limited and no ontology exists with treatments.

DATA

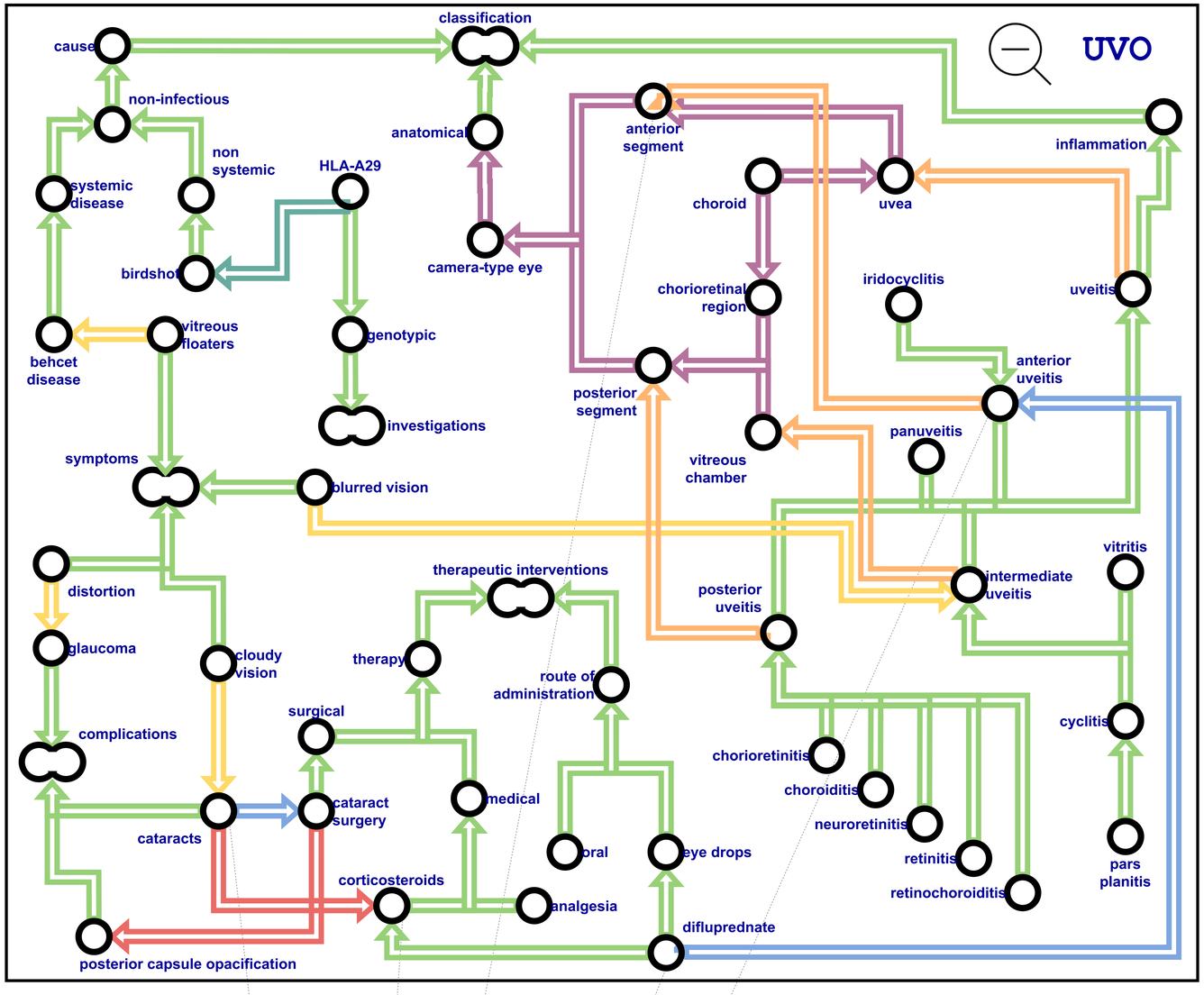
- Medical: Uveitis Clinical Dataset, produced by Uveitis expertise [2].
- Public: Olivia's Vision user forum (oliviasvision.org).

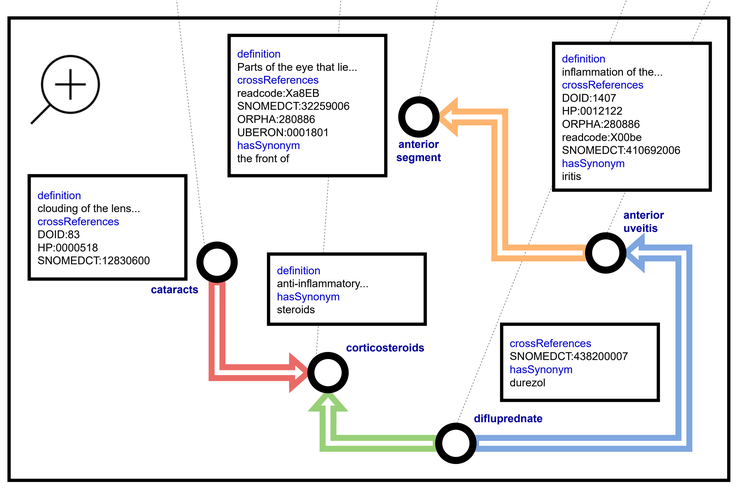
METHODS

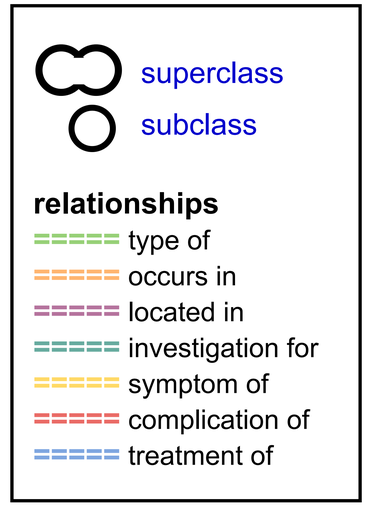
- Foundation of the ontology to be built with the medical data, using Protégé: manual and programmatic construction [3].
- Synonym building with the public data, via text mining with NLTK [4].
- Text Mining Olivia's Vision User Forum
  - 2159 posts (416 threads, treated like cases) from their user forum.
  - Cleaning data (lowercase, removing numbers/special characters, removing stop-words "the/hi/thanks", and stemming which includes "de-pluralise").
  - Equal frequency binning of threads, then split into training / testing set.
  - Term frequency–inverse document frequency (tf-idf).
  - Add important words to ontology, remove words from data, re-run tf-idf for re-weighting.

RESULTS & FUTURE WORK

- Successfully constructed an ontology using medical and public data.
- We need validation: annotating the Olivia's Vision test set to observe if synonyms are evenly distributed.
- Using the ontology for characterising patient cohorts to find underlying relationships.







REFERENCES [1] Gkoutos, G., Schofield, N., and Hoehndorf, R. "The anatomy of phenotype ontologies: principles, properties and applications." *Briefings in Bioinformatics* 19.5 (2017): 1008-1021. [2] Denniston, A. et al. "Uveitis Dataset." *Royal College of Ophthalmologists* (2018). [3] Musen, Mark. "The protégé project: a look back and a look forward." *AI matters* 1.4 (2015): 4. [4] Bird, S., Klein, E., and Loper, E. "Natural language processing with Python: analyzing text with the natural language toolkit." *O'Reilly Media, Inc.* (2009).

@sap218

scp887@student.bham.ac.uk