Profiling Medical Journal Articles Using a Gene Ontology Semantic Tagger

Mahmoud El-Haj
Paul Rayson
Scott Piao
Jo Knight

Origin and Outcomes

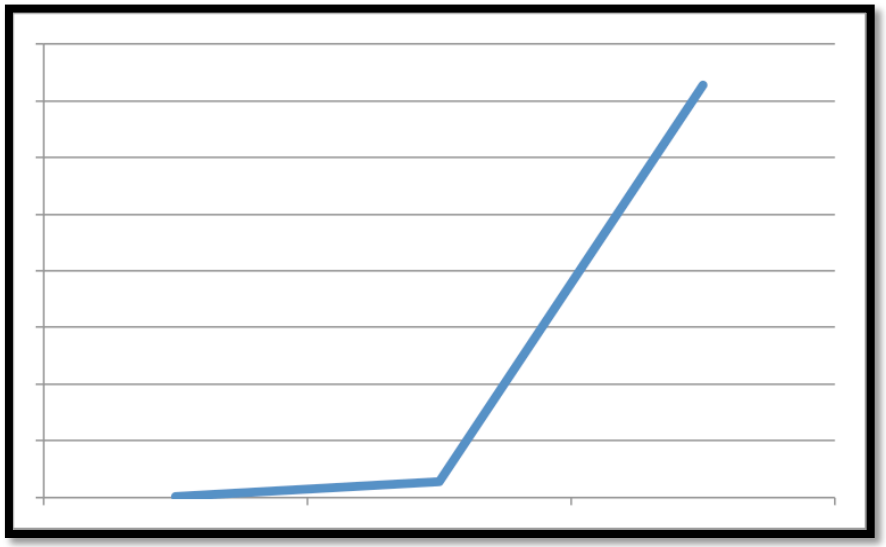
- Currently funded through a Wellcome Trust Seed award
- Collaboration with UCREL through DSI
- International Genetic Epidemiology Society 2017 - Poster presented
- Language Resources Evaluation Conference 2018 - Paper accepted
- Talks Valencia (Paul) /DSI (Jo)
- Future - Section of and EPSRC Grant with Richard Harper ISF

Introduction

- Goal of Human Medical Genetics

Introduction

- Goal of Human Medical Genetics
- Literature explosion



- The need to adapt NLP and Corpus Linguistic methods

Dataset

- Medical journal abstracts from PubMed
- English articles discussing human genetics studies in psychiatry and immune related disorders.

Dataset

Corpus	#Articles	#Words	Keywords
Immune	21.5K	4.8M	(geneti* OR gene OR genot*) AND (immunol* OR immunog* OR immune)
Psychiatric	15.2K	2.8M	(geneti* OR gene OR genot*) AND (psychi)
Reference	296.5K	79.0M	(geneti* OR gene OR genot*)
Total	333.2K	86.7M	

Data Extraction

- Search PubMed website directly
- Saved results to large XML file
- Built a Java Suite for parsing PubMed XML file format.
- Java suite extracts abstracts, titles, authors, pub-date, DOI …etc.
- Code freely available on github:

  https://github.com/drelhaj/BioTextMining

Fine-grained Medical Terms

- Words in pubmed just aren't the same...cytokines, lymphocyte mediated immunity
- Extra level of annotation required for tagging
- The Gene Ontology Consortium's[1] OBO Basic Gene Ontology (go-basic.obo) categories[2].

---

1 http://geneontology.org/

2 http://purl.obolibrary.org/obo/go/go-basic.obo

What is GO?

- **Gene Ontology (GO)** : consistent descriptions of gene products across databases.
- **go-basic.obo**: is the basic version of the GO ontology, filtered such that the graph is guaranteed to be acyclic paths,
- Annotations can be propagated up the graph.
- We focused on the **is_a** relation in order to trace ancestors and children for each entry in the ontology.

Gene Ontology Semantic Tagger (GOST)

- Corpora uploaded to Wmatrix
- POS tagged using CLAWS.
- Semantically tagged using USAS
- Counted frequencies
- Compared sub-corpora using methods from Corpus Linguistics.

Parsing OBO

- we created Java code that combines the use of publicly available OBO library[1]
- with Java Directed Graph (Digraphs)
- to trace the paths from a node child to the root.
- The code used Breadth First and Depth First algorithms to quickly and accurately extract the paths.

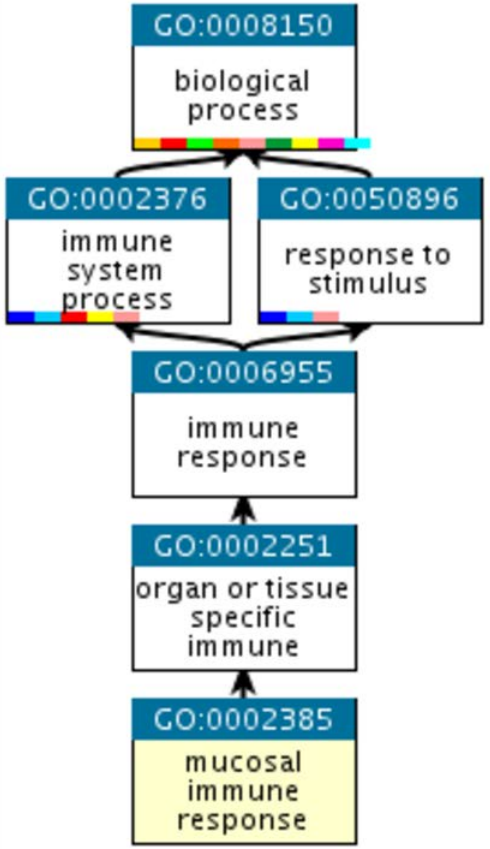
---

1 https://github.com/sugang/bioparser

2 http://purl.obolibrary.org/obo/go/go-basic.obo

OBO Graph Sample

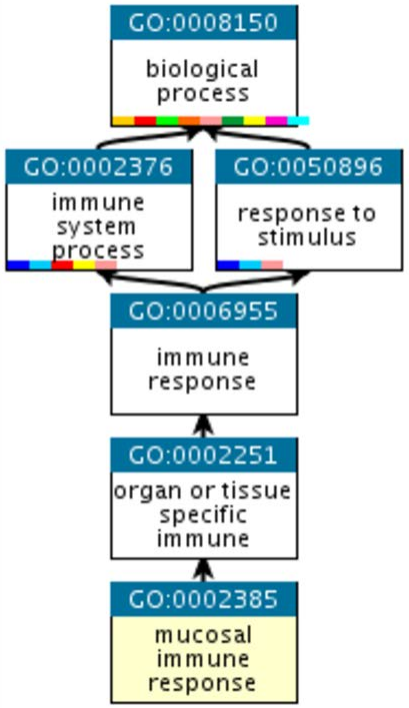
- Our code allowed us to generate a USAS tagger dictionary file
- where each entry in the OBO ontology is tagged with the GO IDs shown in its path.
- In the figure we can see two paths from the child node towards the ``biological process'' root.



Dictionary Creation

The dictionary creation process works as follows:

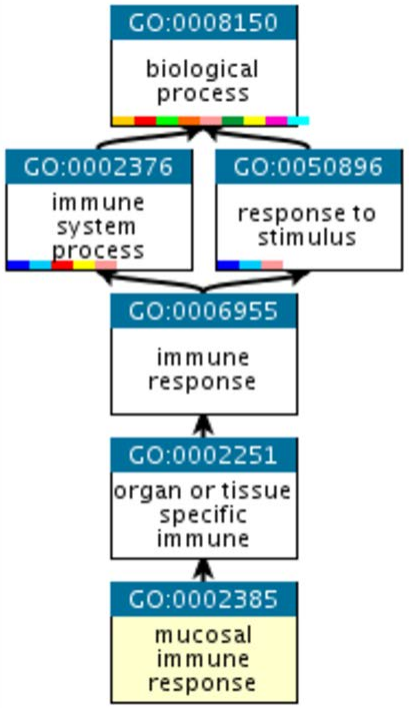
1. Is child node single word or multi-word expression.
2. get number of paths towards the root.
3. get each path's GoID entries (child node's ancestors)
4. include the level of each ancestor by adding that to the end of each entry (e.g. .1 to refer to the first parent (GOO:0002251).
5. Check if path passes through an ``immune system process'' (i.e. GoID: 0002376).
6. If so we add .I to the end of the GoID tag to refer to immune entry, otherwise we add .N referring to a non-immune entry.



Tagging Example

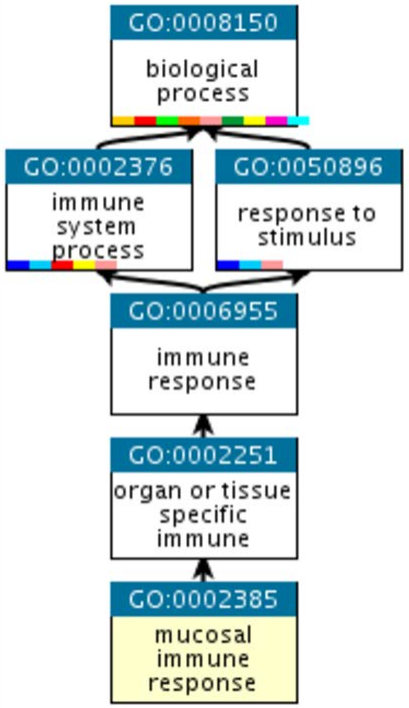
- Following the steps in previous slide, the child node GO:0002385 is multi-word expression entry with following semantic dictionary tags:

- {GO:0008150.4.I, GO:0002376.3.I, GO:0050896.3.N, GO:0006955.2.I, GO:0002385.0.I, GO:0002251.1.N, GO:0006955.2.N, GO:0002385.0.N, GO:0002251.1.I, GO:0008150.4.N}.



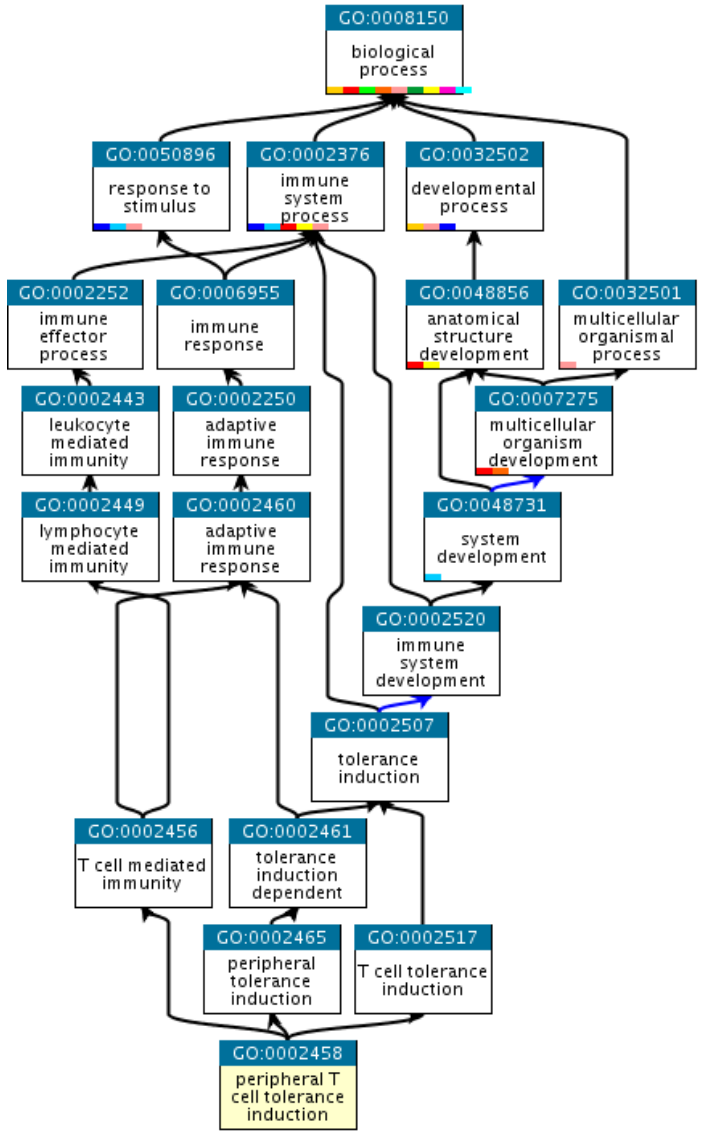
Tagging Example

- Tags such as GO:0006955 ends with .2 suffix referring to level two (counting from level zero).
- and will appear twice;
  - once as an immune entry with a .I suffix (GO:0006955.2.I)
  - and another as a non-immune entry with a .N suffix (GO:0006955.2.N).



Complex Example

- Dictionary creation can be complex
  - Overlapping hierarchies
  - Levels that can be skipped



GOST

- The resultant GO term and ID map collection from the process described above contains:
  - 433 single word bioterms
  - and 44,180 multiword bioterms
- merged into the Lancaster UCREL Semantic lexicons to create a new version of the Lancaster USAS semantic annotation system named:

"GOST" (Gene Ontology Semantic Tagger)

Using The GOST

- Using the GOST, we have tagged 237,615 PubMed abstracts in our corpus.
- This corpus provides a valuable new resource for mining Biomedical and health information from the Biomedical literature.
- The table shows a sample from a tagged abstract, where the part-of-speech tags are from CLAWS C7 tagset
- the generic semantic tags are from the USAS tagset,
- and the MWE tags encode multiword term information including sequential number, term length and location of each word in the given term.

WORD	LEMMA	POS	SEM	MWE
several	several	DA2	N5	0
processes	process	NN2	A1.1.1 X4.2	0
potentially	potentially	RR	A7+	0
involved	involved	JJ	A1.8+ A12-	0
in	in	II	Z5	0
MN	mn	FO	Z99	0
,	PUNC	YCOM	PUNC	0
including	including	II	A1.8+	0
extracellular	extracellular	JJ	GO:0022617.0.N	1:3:1
matrix	matrix	NN1	GO:0022617.0.N	1:3:2
disassembly	disassembly	RR	GO:0022617.0.N	1:3:3
and	and	CC	Z5	0
organization	organization	NN1	S5+c S7.1+	0
,	PUNC	YCOM	PUNC	0
cell	cell	NN1	GO:0007155.0.N	2:2:1
adhesion	adhesion	NN1	GO:0007155.0.N	2:2:2
,	PUNC	YCOM	PUNC	0
cell-cell	cell-cell	JJ	Z99	0
signaling	signaling	NN1	GO:0023052.0.N	0
,	PUNC	YCOM	PUNC	0
cellular	cellular	JJ	GO:0044267.0.N	3:4:1
protein	protein	NN1	GO:0044267.0.N	3:4:2
metabolic	metabolic	JJ	GO:0044267.0.N	3:4:3
process	process	NN1	GO:0044267.0.N	3:4:4
,	PUNC	YCOM	PUNC	0

Results – word comparison

Results – word comparison next level down

- Less predictable words such as "risk''
  - Language is used different despite both corpora describing genetic studies of a complex trait

Results - new GOST annotated corpora

GOID	Name	Immune	%	Psych	%	O/U	Keyness
GO:0005623	cell	33346	7.31	1524	1.02	+	10696.95
GO:0005575	Cellular Component	34577	7.58	1808	1.20	+	10332.02
GO:0007610	behavior	199	0.04	2095	1.40	-	4611.01
GO:0032501	multicellular organismal process	616	0.13	2364	1.57	-	3915.62
GO:0002376	immune system process	7253	1.59	88	0.06	+	3416.63
GO:0008150	Biological Process	7253	1.59	88	0.06	+	3416.63
GO:0006955	immune response	6992	1.53	84	0.06	+	3298.74
GO:0006955	immune response	6992	1.53	84	0.06	+	3298.74
GO:0050877	neurological system process	426	0.09	1756	1.17	-	2991.92
GO:0050896	response to stimulus	7034	1.54	192	0.13	+	2764.12
GO:0002376	immune system process	2958	0.65	28	0.02	+	1443.03
GO:0008150	Biological Process	2933	0.64	28	0.02	+	1429.29
GO:0050890	cognition	10	0.00	536	0.36	-	1402.85
GO:0050877	neurological system process	16	0.00	548	0.37	-	1394.05
GO:0005575	Cellular Component	5013	1.10	308	0.21	+	1357.84

Conclusion and Future Work

- A method for the creation of a semantic lexicon from an existing Gene Ontology, a Gene Ontology Semantic Tagger (GOST)
- Applied to corpora of scientific papers
- Provided a freely available annotated corpora
- Demonstrated the tools extending corpus and computational linguistics allows genomics researchers to get sensible answers

Resources

- The corpora and Java code to parse and annotate the dataset in addition to the ontology lexicon are made publicly available for research purposes.

  https://github.com/drelhaj/BioTextMining

- The Gene Ontology Semantic Tagger will soon be released via the downloadable graphical interface.

  http://ucrel.lancs.ac.uk/usas/gui/

- Project information

  http://wp.lancs.ac.uk/btm/