Profiling Medical Journal Articles Using a Gene Ontology Semantic Tagger

Mahmoud El-Haj[1], **Paul Rayson**[1], **Scott Piao**[1] **and Jo Knight**[2]
[1]School of Computing and Communications and [2]Medical School, Lancaster University, UK
{m.el-haj, p.rayson, s.piao, jo.knight}@lancaster.ac.uk

Abstract

In many areas of academic publishing, there is an explosion of literature, and sub-division of fields into subfields, leading to stove-piping where sub-communities of expertise become disconnected from each other. This is especially true in the genetics literature over the last 10 years where researchers are no longer able to maintain knowledge of previously related areas. This paper extends several approaches based on natural language processing and corpus linguistics which allow us to examine corpora derived from bodies of genetics literature and will help to make comparisons and improve retrieval methods using domain knowledge via an existing gene ontology. We derived two open access medical journal corpora from PubMed related to psychiatric genetics and immune disorder genetics. We created a novel Gene Ontology Semantic Tagger (GOST) and lexicon to annotate the corpora and are then able to compare subsets of literature to understand the relative distributions of genetic terminology, thereby enabling researchers to make improved connections between them.

Keywords: semantic tagger, ontology, genetics, medical

1. Introduction

The explosion of scientific literature in all fields makes it hard to keep apace of new knowledge. This is particularly true in the relatively new field of genomics. For example, a search in the main citation database for biomedical literature (PubMed) for the term 'genome wide association study' results in just 5 papers from 1995, 141 from 2005 and 3,633 from 2015. We contend that the myriad of techniques developed in Information Retrieval coupled with Natural Language Processing can help address these scaling and searching issues. Such a set of techniques could help in a myriad of ways, for example, summarisation of papers or a set of papers, collocation methods to investigate drug-disease-gene interactions, and query expansion where terminology varies from one subfield to another. Previously such techniques have been used to perform tasks such as identifying gene-gene or gene-phenotype interactions (Bundschus et al., 2008; Kann, 2007). In addition, by using corpus comparison methods originating in Corpus Linguistics, we aim to identify key words and concepts emerging from a body of literature that will provide new clues to disease aetiology.

In the remainder of this paper, we describe related work on biomedical text mining and corpus comparison. Then we explain how we created an open access corpus derived from medical journal abstracts, and a novel semantic tagger to apply a lexicon derived from a standard Gene Ontology. Finally, we illustrate how these new resources allow us to profile medical journal articles using domain-specific ontologies.

2. Related Work

Over a number of years, Natural Language Processing (NLP) techniques have been widely applied to biomedical text mining to facilitate large-scale information extraction and knowledge discovery from the rapidly increasing body of biomedical literature. Substantial efforts have been dedicated to this research area. Among the early researchers in this area are Ananiadou et al. (2006), who identified the challenging issue of finding useful information from the plethora of biomedical scientific literature which are manually unmanageable. Kann (2007) also suggested that Text Mining approaches are essential for discovering information about disease and protein interactions buried within millions of biomedical records. Since the recognition of the importance of the Biomedical text mining, a variety of NLP tools have been developed and modified to support it. Among the main tools and corpora developed for such purposes include Genia tagger/corpus (Tsuruoka et al., 2005; Thompson et al., 2017), Termine[1], and LAPPS GRID (Ide et al., 2016). These tools have typically focused only on lexical, syntactic and shallow semantic (named-entity) approaches. Another related biomedical annotation tool is the Penn BioTagger[2] (Jin et al., 2006), which is capable of tagging gene entities, genomic variations entities and malignancy type entities. Despite the progress over the past years, there are still various issues which remain unsolved, including the lack of NLP tools tailored for specific subfields of biomedical research, and the need to link entities at the conceptual level. In this work, we report on our experiment in which we modify a semantic tagger and create a corpus semantically tagged with both related sub-sets of the Gene Ontology categories and generic semantic field categories for an aetiology study.

Comparing corpora is a key method in corpus linguistics, and is a vital step towards measuring the differences between collections of textual documents. Previous approaches have been focused on word level comparisons only, finding terms or keywords that can differentiate one corpus from the other (Kilgarriff, 2001; Rayson and Garside, 2000). When the method is applied at the semantic level (for example with the general purpose USAS taxonomy[3]), this enables confirmation of the word-level findings but also the ability to uncover key semantic categories, which are more dispersed across a wider group of words and would not otherwise be highlighted as key (Rayson, 2008). In a medical context, we hypothesise that it is impor-

[1]http://www.nactem.ac.uk/software/termine/
[2]http://seas.upenn.edu/∼strctlrn/BioTagger/BioTagger.html
[3]http://ucrel.lancs.ac.uk/usas/

tant to use a more fine-grained semantic taxonomy which embodies greater medical domain knowledge, hence our undertaking the research presented here which derives and applies a gene ontology semantic lexicon to this problem.

3. Dataset

We collected medical journal abstracts from PubMed[4] by restricting the search to retrieve only English medical articles discussing human genetics studies in psychiatry and immune related disorders. Table 1 shows the dataset statistics in terms of article and word counts. The searches have been adapted to ensure appropriate literature coverage. For example, whilst including `immun*` in the abstract search picks up papers on many diseases such as psoriasis, the same approach using the term `psych*` is not as effective. In our results, we directly compare the Immune and Psychiatric subcorpora only, but the Reference dataset statistics are included here to show the relative size of the two subcorpora. We will also be employing the Reference corpus in other experiments and to check vocabulary coverage of the existing semantic lexicon. We chose immune and psychiatric genetics corpora as examples that would be very different from each other allowing us to test the utility of the tools. The selected domains fall within the fourth author's research expertise and this has helped in appropriately interpreting the findings (Pouget et al., 2016).

The dataset was downloaded from PubMed in large XML file format[5]. We built a Java suite for parsing PubMed XML file format and extract abstracts along with other information such as journal titles, author names, publication date, DOI and so on. Our code is publicly available for research purposes.[6]

Table 1: Corpus Statistics

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	

4. Gene Ontology Semantic Tagger

For our initial experiments, the corpora were uploaded to Wmatrix[7] where we ran automatic part-of-speech tagging using CLAWS, semantic field tagging using USAS, and counted word frequencies and compared sub-corpora using the keywords method from corpus linguistics. We quickly realised that we needed to provide better coverage of the more fine-grained medical terminology in the PubMed corpora, and therefore included an extra level of annotation by tagging the corpora using The Gene Ontology Consortium's[8] OBO Basic Gene Ontology (go-basic.obo) categories[9].

The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products across databases. The *go-basic.obo* is the basic version of the GO ontology, filtered such that the graph is guaranteed to be acyclic paths, and 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. We chose the *is_a* relationship in the first instance because it has a more intuitive meaning. Something is only considered *is_a* if an instance of the child process is an instance of the entire parent process.

To parse the OBO file we created Java code that combines the use of publicly available OBO library[10] 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. Figure 1 shows an example of a directed graph for the *basophil homeostasis* GO entry. The figure shows two paths starting from the child entry up to the biological process root.

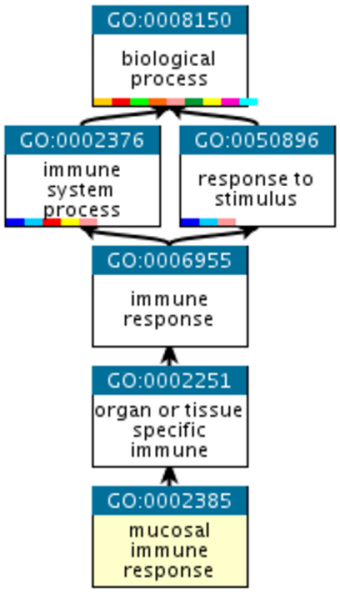

Figure 1: GO Directed Graph Sample

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. Taking the "mucosal immune response" OBO entry shown in Figure 1 we can see there are two paths starting from the child node towards the "biological process" root. The dictionary creation process works as follows:

[4] https://www.ncbi.nlm.nih.gov/pubmed

[5] Instead of using PubMed API we searched PubMed website directly and exported the results to XML using PubMed "Send To File" service.

[6] https://github.com/drelhaj/BioTextMining

[7] http://ucrel.lancs.ac.uk/wmatrix/

[8] http://geneontology.org/

[9] http://purl.obolibrary.org/obo/go/go-basic.obo

[10] https://github.com/sugang/bioparser

1. determine whether the child node is single word or multi-word expression. The example shows the latter.
2. determine the 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. determine whether the path passes through an "immune system process", which is the one with GoID: 0002376. 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.

Following the steps above, the child node GO:0002385 will be considered a multi-word expression entry and will have the 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.`
In the above dictionary, tags such as GO:0006955 will be extended with a `.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). While the GO directed graph snippet shown in figure 1 is relatively simple, figure 2 shows a much more complex example illustrating that the dictionary creation process can become more troublesome with overlapping hierarchies and levels that can be skipped for some graph traversals.
The resultant GO term and ID map collection from the process described above, which contains 433 single word bioterms and 44,180 multiword bioterms, has been merged into the Lancaster UCREL Semantic lexicons to create a new version of the Lancaster USAS semantic annotation system (Rayson et al., 2004; Piao et al., 2017), named GOST (Gene Ontology Semantic Tagger), in order to automatically annotate the bioterms with GO IDs in the journal articles, along with generic USAS semantic tags. Currently, 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.
Table 2 shows a sample from a tagged abstract, where the part-of-speech tags are from CLAWS C7 tagset[11], the generic semantic tags are from the USAS tagset[12], the tags with leading code *GO* are from the Gene Ontology, and the MWE tags encode multiword term information including sequential number, term length and location of each word in the given term. As shown in the table, such a tagging can facilitate analysis of Gene information at any hierarchical levels of the Gene Ontology as researchers need. For example, researchers can

[11]http://ucrel.lancs.ac.uk/claws7tags.html
[12]http://ucrel.lancs.ac.uk/usas/

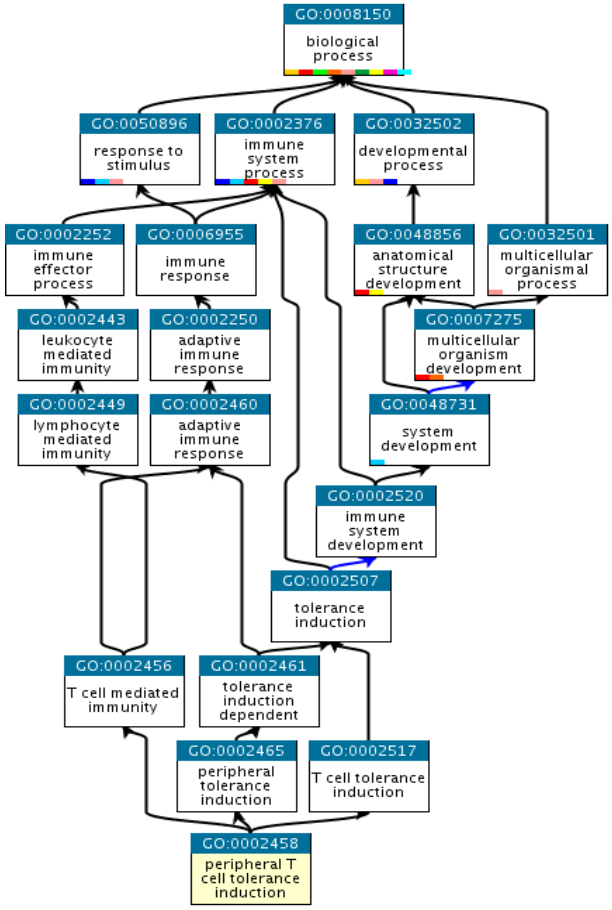

Figure 2: GO Directed Graph More Complex Sample

filter their analysis results by setting a range of hierarchical levels of [3-4], in which case only GO categories {GO:0008152.3.N, GO:0071704.3.N, GO:0008150.4.N, GO:0009987.3.N, GO:0006807.3.N, GO:0008152.4.N} would be considered for the term "cellular protein metabolic process" in Table 2.

5. Results

In our preliminary work using only a word level comparison (El-Haj et al., 2017), we uncovered many subject specific words have a much higher proportional representation in one corpus (e.g. schizophrenia). Other less predictable words such as "risk" are also found to be more frequent in psychiatric literature. The increased proportional representation suggests that language is used different despite both corpora describing genetic studies of a complex trait.
With the new GOST annotated corpora, we are able to compare the two corpora at the semantic level using the Gene Ontology concepts, see Table 3 for keyness sorted results. The final six columns show the actual and relative frequencies for the immune and psych sub-corpora, an indication of over- and under-use (by a direct comparison of the relative frequencies) and the log-likelihood keyness value. Many of the GO terms with the most significantly different frequencies between the two corpora are those strongly related to the suspected biological underpinning of the traits. For example "immune response", "immune system process" and "response to stimulus" were all more frequent in the im-

Table 2: Sample tagged text

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 GO:0016043.3.N GO:0044763.2.N GO:0043062.2.N GO:0030198.1.N GO:0016043.2.N GO:0008150.4.N GO:0044699.3.N GO:0022411.1.N GO:0044763.3.N GO:0044699.4.N GO:0071840.4.N GO:0071840.3.N GO:0022617.0.N GO:0009987.3.N GO:0008150.5.N GO:0009987.4.N	1:3:1
matrix	matrix	NN1	GO:0022617.0.N GO:0016043.3.N GO:0044763.2.N GO:0043062.2.N GO:0030198.1.N GO:0016043.2.N GO:0008150.4.N GO:0044699.3.N GO:0022411.1.N GO:0044763.3.N GO:0044699.4.N GO:0071840.4.N GO:0071840.3.N GO:0022617.0.N GO:0009987.3.N GO:0008150.5.N GO:0009987.4.N	1:3:2
disassembly	disassembly	RR	GO:0022617.0.N GO:0016043.3.N GO:0044763.2.N GO:0043062.2.N GO:0030198.1.N GO:0016043.2.N GO:0008150.4.N GO:0044699.3.N GO:0022411.1.N GO:0044763.3.N GO:0044699.4.N GO:0071840.4.N GO:0071840.3.N GO:0022617.0.N GO:0009987.3.N GO:0008150.5.N GO:0009987.4.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 GO:0022610.1.N GO:0008150.2.N	2:2:1
adhesion	adhesion	NN1	GO:0007155.0.N GO:0022610.1.N GO:0008150.2.N	2:2:2
,	PUNC	YCOM	PUNC	0
cell-cell	cell-cell	JJ	Z99	0
signaling	signaling	NN1	GO:0023052.0.N GO:0008150.1.N	0
,	PUNC	YCOM	PUNC	0
cellular	cellular	JJ	GO:0008152.3.N GO:0019538.1.N GO:1901564.2.N GO:0071704.3.N GO:0044267.0.N GO:0008150.4.N GO:0044260.1.N GO:0044237.2.N GO:0043170.2.N GO:0044238.2.N GO:0009987.3.N GO:0006807.3.N GO:0008150.5.N GO:0008152.4.N	3:4:1
protein	protein	NN1	GO:0008152.3.N GO:0019538.1.N GO:1901564.2.N GO:0071704.3.N GO:0044267.0.N GO:0008150.4.N GO:0044260.1.N GO:0044237.2.N GO:0043170.2.N GO:0044238.2.N GO:0009987.3.N GO:0006807.3.N GO:0008150.5.N GO:0008152.4.N	3:4:2
metabolic	metabolic	JJ	GO:0008152.3.N GO:0019538.1.N GO:1901564.2.N GO:0071704.3.N GO:0044267.0.N GO:0008150.4.N GO:0044260.1.N GO:0044237.2.N GO:0043170.2.N GO:0044238.2.N GO:0009987.3.N GO:0006807.3.N GO:0008150.5.N GO:0008152.4.N	3:4:3
process	process	NN1	GO:0008152.3.N GO:0019538.1.N GO:1901564.2.N GO:0071704.3.N GO:0044267.0.N GO:0008150.4.N GO:0044260.1.N GO:0044237.2.N GO:0043170.2.N GO:0044238.2.N GO:0009987.3.N GO:0006807.3.N GO:0008150.5.N GO:0008152.4.N	3:4:4
,	PUNC	YCOM	PUNC	0

Table 3: Gene Ontology Semantic Keyness Results

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

mune disorder related corpus. The following terms were more frequent in the psychiatric corpus ("neurological system process" and "cognition"). Some of these terms are expected, and help to confirm that our methodology is successful and some categories offer routes for further investigation. We have therefore proved that in principal the method is working and we will continue to mine the results for biological insight.

6. Conclusion and Future Work

In this paper, we have illustrated our early explorations into extending corpus and computational linguistics methods to permit genomics researchers to explore their rapidly growing literature in new ways. Our main contributions are the corpus-based explorations of the research literature on human genetics studies, a method for the creation of a semantic lexicon from an existing Gene Ontology, a Gene Ontology Semantic Tagger (GOST) to apply this to corpora of scientific papers, and freely available annotated corpora. In terms of future work, we will further investigate how our new fine-grained taxonomy performs in terms of contextual accuracy, and whether the level of detail introduced is too much for our planned application. We have already investigated this type of fine-grained task in research related to historical contexts with the Historical Thesaurus Semantic Tagger (Piao et al., 2017). Here, there may need to be a compromise between levels of accuracy and domain-specific explainability. We also intend to carry out a corpus-based investigation into variability of GO terms which may not be replicated exactly in the corpora, for example inflectional or derivational suffixes such as "processes" instead of "process", and the potential for intervening items within multiword expressions. This will allow us to increase the

tagger's accuracy as well as potentially offering a semi-automatic route for updating GO itself. The corpora and Java code to parse and annotate the dataset in addition to the ontology lexicon are made publicly available for research purposes.[13] The Gene Ontology Semantic Tagger has also been released via the downloadable graphical interface[14].

7. Acknowledgements

We are grateful for support for this research from a Wellcome Trust Seed funding grant (reference 204475/Z/16/Z). For more details about our research, see the project website: http://wp.lancs.ac.uk/btm/

8. Bibliographical References

Ananiadou, S., Kell, D. B., and Tsujii, J. (2006). Text mining and its potential applications in systems biology. *Trends in Biotechnology*, 24(12):571–579.

Bundschus, M., Dejori, M., Stetter, M., Tresp, V., and Kriegel, H.-P. (2008). Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinformatics*, 9(1):207, Apr.

El-Haj, M., Piao, S., Rayson, P., and Knight, J. (2017). A comparison between genetics papers relating to immune disorders and psychiatric disorders. In *Proceedings of The 2017 Annual Meeting of the International Genetic Epidemiology Society*.

Ide, N., Suderman, K., Pustejovsky, J., Verhagen, M., and Cieri, C. (2016). The language application grid and galaxy. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Jin, Y., McDonald, R. T., Lerman, K., Mandel, M. A., Carroll, S., Liberman, M. Y., Pereira, F. C., Winters, R. S., and White, P. S. (2006). Automated recognition of malignancy mentions in biomedical literature. *BMC Bioinformatics*, 7(492).

Kann, M. G. (2007). Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Briefings in Bioinformatics*, 8(5):333–346.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

Piao, S., Dallachy, F., Baron, A., Demmen, J., Wattam, S., Durkin, P., McCracken, J., Rayson, P., and Alexander, M. (2017). Time-sensitive historical thesaurus-based semantic tagger for deep semantic annotation. *Computer Speech and Language*, 46:113–135.

Pouget, J. G., GonÃ$\mathbf{S}_a$lves, V. F., , Spain, S. L., Finucane, H. K., Raychaudhuri, S., Kennedy, J. L., and Knight, J. (2016). Genome-wide association studies suggest limited immune gene enrichment in schizophrenia compared to 5 autoimmune diseases. *Schizophrenia Bulletin*, 42(5):1176–1184.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora - Volume 9*, WCC'00, pages 1–6, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rayson, P., Archer, D., Piao, S., and McEnery, A. (2004). The ucrel semantic analysis system. In *Proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop*, pages 7–12.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4):519–549.

Thompson, P., Ananiadou, S., and Tsujii, J., (2017). *Handbook of Linguistic Annotation*, chapter The GENIA Corpus: Annotation Levels and Applications. Springer, Dordrecht, Netherlands.

Tsuruoka, Y., Tateishi, Y., Kim, J.-D., Ohta, T., McNaught, J., Ananiadou, S., and Tsujii, J. (2005). Developing a robust part-of-speech tagger for biomedical text. In *Advances in Informatics - 10th Panhellenic Conference on Informatics (PCI 2005), LNCS 3746*, pages 382–392.

[13] https://github.com/drelhaj/BioTextMining

[14] http://ucrel.lancs.ac.uk/usas/gui/