HG2BTM Workshop: Hypothesis Generating in Genetics and Biomedical Text Mining
Workshop Programme: 8th January 2019

10.00 - 10.45: Talk 1 - Dr Paul Thompson, Centre for Corpus Research, University of Birmingham, UK: *Corpus linguistic insights on specialised discourses*

10.45 - 11.30: Talk 2 - Lancaster University Team: *An introduction to the BioTM project*

11.30 - 11.45: Morning coffee break

11.45 - 12.30: Talk 3 - Professor Nancy Ide, Department of Computer Science, Vassar College, USA: *An Environment for Biomedical Text Mining: The LAPPS Grid*

12.30 - 13.15. Working lunch and discussion

13.15 - 13.30: Comfort break

13.30 - 14.15: Talk 4 - Dr Mark Stevenson, Department of Computer Science, Sheffield University, UK: *Discovering Hidden Knowledge from Scientific Literature: Challenges and Some Solutions*

14.15 - 15.00: Talk 5 - Professor Udo Hahn, Jena University Language & Information Engineering (JULIE) Lab at University of Jena, Germany: *Gene Interaction Hypothesis Generation Support: Tooling Technology Biologists Really Like*

15.00 - 16.00: Discussion, future directions, and collaborative possibilities over afternoon coffee

Abstracts

Professor Nancy Ide: *An Environment for Biomedical Text Mining: The LAPPS Grid*:

Abstract: It is widely recognized that the ability to exploit Natural Language Processing (NLP) text mining strategies has the potential to increase productivity and innovation in the sciences by orders of magnitude. However, to date the use of NLP technologies has required considerable skill in the field. The Language Applications (LAPPS) Grid (Ide et al., 2014) provides an infrastructure for rapid development of natural language processing applications (NLP) by providing access to a wide range of tools and making them both syntactically and semantically interoperable. As such, the LAPPS Grid provides an intuitive and easy-to-use platform that enables users to experiment with and exploit NLP tools and resources without the need to determine which are suited to a particular task, and without the need for significant computer expertise. With funding from the US National Science Foundation, we are adapting the LAPPS Grid to serve the needs of scientists who wish to search and mine vast bodies of scientific publications.

Currently, an instance of the LAPPS Grid tailored to mining biomedical publications is currently maintained on the JetStream cloud environment, which currently includes enables access to PubMed (ca. 12 million abstracts) and PubMed Central (over 11 thousand full-text documents) as well as full-text PubMed data used in several BioNLP and SemEval shared tasks. In addition, the LAPPS Grid has collaborated with the developers of PubAnnotation to integrate the services and resources provided by each in order to greatly enhance the user's ability to annotate scientific publications and share the results. LAPPS Grid users can access annotations in the PubAnnotation repository that are linked to the texts and add or edit annotations using PubAnnotation's TextAE, a powerful and easy-to-use Javascript app for text annotation and visualization. Reciprocal access between PubAnnotation and the LAPPS Grid means that users can easily apply automatic annotation tools and subsequently manually correct annotations, in an iterative "human-in-the- loop" process of refinement. This, coupled with the LAPPS Grid state-of-the-art Open Advancement evaluation facilities, provides a powerful environment for rapid development of high-quality automatic annotation procedures that can develop training data for machine learning.

Professor Udo Hahn: Exploring Genes and their Interactions: Text Analytics for the Needs of Biologists

In this talk, I will focus on different ways natural language processing (NLP) technology can support life scientists to maneuver in the document and knowledge spaces made up by an already vast and ever-increasing number of publications and internal reports. These documents encode myriads of knowledge, yet in an unstructured way (from the perspective of computers, at least). The basic challenge for NLP technology is to transform the contents of these documents into a format that is easier and faster accessible for content-based computational processing (querying and retrieving, browsing, zooming, summarizing, etc.). The methodological approaches underlying this task can be divided into three fundamental information service classes – document retrieval, fact retrieval and knowledge discovery. I will discuss the main distinctions and interdependencies underlying these services and illustrate them with systems (mostly) developed in the Jena University Language & Information Engineering (JULIE) Lab at the University of Jena.

Dr Mark Stevenson Discovering Hidden Knowledge from Scientific Literature: Challenges and Some Solutions

Hidden knowledge occurs when a connection can be inferred by combining information from multiple documents but that connection has not been noticed. Literature based discovery (LBD) aims to identify hidden knowledge and has been applied to the scientific literature with some notable successes. This talk discusses some of the challenges that are faced when applying LBD to large collections of scientific literature, particularly ones related to volume, ambiguity and inconsistency. It will also present some approaches that can be used to alleviate these problems.