Annex No. 10 to the MU Directive on Habilitation Procedures and Professor Appointment Procedures

HABILITATION THESIS REVIEWER'S REPORT

Masaryk University	
Applicant	Vít Nováček
Habilitation thesis	A Journey in Biomedical Discovery Informatics: From Ontology Learning to Knowledge Graph Embeddings
Reviewer	Prof. Paul Schofield
Reviewer's home unit, institution	University of Cambridge, United Kingdom

Dr Novacek has submitted a series of thematically linked papers; six primary research papers in peer reviewed journals, one review and two papers in conference proceedings. The main theme of the first set of papers is the capture and management of knowledge and the development in the second theme is of the representation of that knowledge in knowledge graphs and the use of vector representations via embeddings in making this symbolic knowledge amenable to machine learning approaches such as ANN. This is a coherent body of work and spans 2008 to 2020, the first paper being developed from his Master's thesis work. All of the published papers are collaborations with other investigators, some of these in industry and it is very clear that Dr Novacek is the prime or joint prime investigator on these papers. It is possible to see though time his establishing of his own independence and developing his own individual contribution to the discipline..

Papers 1-3 represent approaches to the discovery, extraction and formalisation of concepts from literature and draw heavily on natural language processing. This first paper addresses the problem of large-scale ontology curation and the ways in which automated processes can help with expert curation, especially during continuous updating and take some of the work out of integrating new knowledge into existing ontologies. The motivation was at least in part the problem of lack of high level computer science abilities in many domain experts and the aim was therefore to make the lifecycle of an ontology much more amenable to community management. At the time this work was being conceptualised and executed this was a known problem in the biomedical community and although some work had been done to harvest ontologies directly from knowledge sources such as the published literature I don't think I know of any examples where the various processes preseneted here were integrated into a platform for supporting the complete life cycle of an ontology. This was consequently an ambitious an original study. Unfortunately I feel that in this case uptake and assessment in a real world application doesn't seem to have been attempted and I'm not aware of any major ontology that uses the DINO platform for its development and maintenance. Moreover the implementation of the Ontology Development Kit (ODK) for OBO ontologies in the years since this paper was published have obviated any attempt to make ontology development more accessible to non-computer scientists because of its extreme complexity. This does not however detract from the interest and novelty of the work, and it shows that the applicant was working at the cutting edge in the mid 2000s.

The next two papers focus on semantic approaches to content in published papers, the former with a semantic approach to literature searching and the latter with extraction of knowledge from individual papers in order to provide a rapid – "skim" – reading facility for expert or non-expert readers. At the time the former paper was certainly quite novel and, although limited to certain semantic properties of concept relations within papers, was nevertheless successful and won the Elsevier Grand Challenge prize which was a major achievement. The approach taken in the second paper was very novel and at least partly successful in that they did provide evidence from expert evaluation that concept and relation extraction from the literature could condense a paper in a meaningful way to be skim read by a system user. The paper is rather interesting in its evaluation of the effect of parameters such as pruning on the meaningfulness and utility of the graphs extracted and although again I dont think that this tool is widely used by the general community the approach and the work that went into assessing it are certainly very useful and interesting.

The following thematic papers are amongst the most significant offered, with papers 5 7 and 8 the most novel and useful contributions. This set of papers examines the use of knowledge graphs and particularly embeddings in several different contexts. Paper 4 looks at the effects of using background knowledge to improve the learning of embeddings by neural networks and particularly the augmentation of axioms by equivalence and partonomic relations with significant improvement in the predictive accuracy of several NN embedding models. This is a rather general and fundamental exploration of some of the methodology for learning embeddings and is an interesting but incremental contribution. Much more significant is paper 5 which looks at the problem of predicting adverse drug reactions, and they evaluate different multilabel learning models against several datasets in terms of various multi-label ranking evaluation metrics. This is a comprehensive benchmarking exercise and concludes that off the shelf methods seem to perform better than existing ADR prediction methods. This rather technical paper is a useful addition to the literature and explores the utility of a wide range of methods. It is a shame that the promised work in the influence of embeddings, as mentioned in the conclusions, does not seem to have been done with these datasets.

Paper 6 addresses the import and integration of relational data from public sources such as STRING and CTD into knowledge graphs and presents a pipeline for parsing processing and mapping this data into a set of KG triples making up the BIOkg knowledge graph. This is an interesting resource but the paper is somewhat preliminary and does not contain good data on the behaviour of the pipleline and the resulting accuracy and completeness of the imported data, for example the accuracy of mapping of IDs etc. More information would have been much appreciated but this was a conference proceeding paper and Im sure constrained in size as a consequence.

I consider paper 7 to be one of the most significant. This looks at drug target prediction as a link prediction problem on a knowledge graph using learned embeddings and vector representation. This at the time was becoming a standard general approach but the paper uses a novel combination of tools including the Tri-model KG embedding model, and 10x cross validation on the resulting predictions. This is compared with existing state of the art models and found to outperform them all. Moreover, there was some manual validation of some individual predictions which looked very promising. It will be interesting to see how his work develops.

Paper 9 takes a similar general approach to link prediction over KGs but for kinase-substrate data, which is of great biological interest. This is a very novel approach and combines the generation of a trained KG derived from existing phosphorylation data with a KG of candidate kinase substrate interactions outputting predicted kinase substrate relations not present in the original KG. They benchmark this against existing approaches and most importantly carry out some experimental validation. This is the most important paper of the set and is an important contribution to the field.

The final paper is a well put together critical commentary and an interesting and useful addition.

Overall this set of papers represents the scientific development of a capable and creative informatician and contains several major contributions to the field, particularly paper 8. My assessment is that this would certainly be a profile appropriate for a tenure decision in the UK or the US Universities with which I am familiar. I can only compare the standard with these as we have no equivalent of the habilitation here.

Reviewer's questions for the habilitation thesis defence (number of questions up to the reviewer)

1. The ROC AUC is one of the most often used metrics for algorithmic success in making predictions. Does he believe that this is a good metric, and has he looked to see how predictions are ranked for example in outputs that seem superficially to have good AUCs?
2. To what extent does he believe that overfitting is a problem in the ML approaches he has used, what is his experience of overfitting and has he investigated this in any of his works? How can overfitting be identified?
3. Most of the knowledge graphs have simple axioms and so far as I can recall most of the ones that he has used or developed only have one axiom type in the graph. Has he looked into the problem of having multiple axiom types, ie multiple relation types withing KGs and what problems could he foresee in generating embeddings over such complex KGs?
4. Have the tools developed in paper 8 been taken up and used or further developed for example in the pharmaceutical industry?
5. Did he try learning over large KG using graph convolutional NN? If so what was his experience?
6. Similarly, does he have an experience or thoughts on using generative adversarial networks in generating embeddings?

Conclusion

The habilitation thesis entitled "*A Journey in Biomedical Discovery Informatics: From Ontology Learning to Knowledge Graph Embeddings*" by Vít Nováček. **fulfils** the requirements expected of a habilitation thesis in the field of Informatics.

[redacted]

Date: 20 April 2022 Signature: Prof. Paul Schofield