ABSTRACT:

Name of Responsible Researcher:	Aidan Hogan
Proposal Title:	Enhancing Neural Question Answering Systems for Knowledge Graphs

Describe the main issues to be addressed: objectives, methodology and expected results. **The maximum length for this section is 1 page** (use letter size format, Verdana font size 10 or similar).

Setting: A variety of open knowledge graphs have been published on the Web in recent years using Semantic Web standards, the most prominent of which – including DBpedia, Wikidata, etc. – describe millions of entities spanning a wide variety of domains and receive in the order of millions of queries per day over the Web. The standard way to query such knowledge graphs is through graph query languages, such as SPARQL. However, many users are not expert in such languages and thus struggle to query these knowledge graphs. Even for users expert in graph query languages, querying knowledge graphs can be challenging due to their diversity, incompleteness, and lack of succinct schema.

A number of systems have been proposed to assist both expert and non-expert users to find they answers they seek over knowledge graphs. A promising – though ambitious – approach to help users in this way is that of *Knowledge Graph Question Answering* (KGQA) in which the user specifies their question in natural language and receives answers directly. Within the area of KGQA, recent approaches have had success applying machine-learning architectures based on deep neural networks (specifically, recurrent and convolutional neural networks). These architectures are used to translate questions into structured queries that can be evaluated directly over the knowledge graph. While the results for KGQA neural-based approaches are promising, the performance of such systems depends heavily on the quality of the training dataset used, which features natural language question–structured query pairs. Performance issues arise for neural-based systems when dealing with questions that involve terms of phrases (the *out-of-vocabulary problem*), or novel query structures (the *out-of-template problem*) not seen during training.

Objectives: The primary objective of this research topic is to investigate techniques and develop resources that advance the state-of-the-art performance for neural-based KGQA systems for complex questions. There are two specific objectives: (1) improve the performance of neural-based KGQA systems for a given training and test dataset by developing techniques to address both the out-of-vocabulary and out-of-template problems; (2) improve the quality of the KGQA datasets available for neural-based approaches by creating a collaborative platform for semi-automatically curating question–query pairs of interest to users (specifically for the Wikidata knowledge graph) that can be used for training and testing KGQA systems, and that includes features for automatically canonicalising and generating further question–query pairs.

Methodology: The project will involve a mix of both experimental and theoretical methodologies. In terms of the experimental methodology, the project will begin by developing and exploring novel techniques to enhance neural-based KGQA systems using existing KGQA benchmark datasets, such as the LC-QuAD 2.0 dataset defined for Wikidata. These novel KGQA approaches will be compared against state-of-the-art approaches from the literature in terms of accepted metrics, such as the BLEU score for translation quality, the accuracy of the queries generated, and the precision/recall/F_1-score of the results generated. The theoretical methodology will largely play a role for the canonicalisation of queries (modulo query-equivalence), which we hypothesise may improve the performance of neural-based KGQA systems by reducing the syntactic variance in the queries used for training; in particular, the problem of deciding query equivalence is known to be undecidable for query languages such as SPARQL, where the goal will thus be to explore restricted fragments of SPARQL for which sound-but-incomplete or sound-and-complete canonicalisation can be performed.

Expected Results: As a result of the project, we expect to publish 2–3 journal papers and 2–3 conference papers describing novel techniques that advance the state-of-the-art for KGQA, aiming for prestigious international venues. We also aim to have an online KGQA system for Wikidata, a novel KGQA dataset for Wikidata, and a collaborative platform on which KGQA question–query pairs for Wikidata can be semi-automatically curated. All code and datasets will be published on the Web under liberal licenses for further extension, re-use, and reproducibility purposes. We expect to financially support 2 Master students and 2 Undergrad students. The project will further involve and partially support a (currently active) PhD student.