Identifying Semantically Similar questions using NLP techniques and Linked Data Principles

Anirban Bhattacharjee, Master of Science in Computer Science

University of Dublin, Trinity College, 2019

Supervisor: Professor Declan O'Sullivan

In Community Question Answering (CQA) sites, despite active participation, a significant amount of questions on such sites remain unanswered due to a lot of reasons such as the question being poorly formed/worded, unavailability of any answerer or the increased inflow of questions in the same area which disinterests an answerer to answer the same question or having to redirect them multiple times to already answered questions. This research is an attempt to study if unstructured data is converted to structured data using state-of-the art natural language processing (NLP) techniques and Linked Data technologies, to what extent it could help a user in identifying semantically similar questions. One of the most contested themes in Computer Science is the ability to automatically map natural language semantics into programming languages. This research work is distinguished from other studies as we approach the problem from an ontology centred view and the idea of knowledge reuse forms the notion of this work. We evaluate our approach and open discussions on new ways to evaluate the identification of semantically similar questions. The key findings of this research demonstrate that using NLP techniques and Linked Data principles identification of semantically similar questions is viable. The proposed approach has a small but significant impact which can be leveraged for designing data models for the task of finding semantically similar questions.