1 Billion Citation Dataset and Deep Learning Citation Extraction

Mark Grennan, Master of Science in Computer Science
University of Dublin, Trinity College, 2019

Supervisor: Joeran Beel

Abstract

Citation or reference parsing involves extracting machine readable metadata from a citation string. This paper details the work carried out in creating a large, diverse and labelled citation dataset which can be used to train ML citation parsing tools. The dataset was created by adapting the citation styles within CSL, collecting citation metadata from CrossRef and using the open-source citation processor, citeproc-js. It contains 991,411,100 XML labelled citation strings in over 1,500 different citation styles. These are composed of a diverse range of citation types from a wide range of academic fields. When the granularity of the training data is the same the 1 Billion Citation Dataset was shown to be on a par with the state-of-the-art hand-labelled training datasets (Hand-labelled Training Data Macro-Average F1 0.74 vs Synthetic Training Data Macro-Average F1 0.74). These results indicate that the 1 Billion Citation Dataset has a lot of potential for training a deep-learning based citation parsing model and that this may lead to significant improvements in the accuracy of citation parsing.