A text mining pipeline for curating information in computational neuroscience using active and deep learning.

Shawn Kruger*

Editorial Office, Journal of Neuroinformatics and Neuroimaging, London, United Kingdom

Accepted May 06, 2021

Commentary

Curation of neuroscience things is critical to current work in neuroinformatics and computational neuroscience, such as those used in large-scale brain modelling projects. Manually combing through hundreds of articles in search of fresh information on modelled entities, on the other hand, is a time-consuming and low-reward activity. Text mining can aid a curator in extracting important data from this material in a methodical manner. The use of text mining algorithms to the neuroscience literature is proposed. Two computational neuroscientists used active learning techniques to annotate a corpus of items relevant to neuroscience, allowing for quick, focused annotation. After that, we trained machine learning models to recognise the specified items. Neuron Types, Brain Regions, Experimental Values, Units, Ion Currents, Channels, and Conductances, as well as Model Organisms, are all discussed. We compared a standard rule-based method, a conditional random field, and a deep learning-based entity recognition model, and found that the deep learning model outperformed the others. With a macro average accuracy, recall, and F1 score of 0.866, 0.817, and 0.837, respectively, we can recognise a variety of named items of interest to neuroscientists. The following are the contributions of this work:

- Provide a collection of Named Entity Recognition (NER) algorithms capable of recognising neurological entities at a level of performance that is superior to or comparable to previous work
- Offering a training approach for NER tools in neuroscience that requires relatively minimal training data to achieve high performance. This method may be used to any sub-domain of neuroscience

Giving a small corpus containing annotations for a variety of object types, as well as annotation guidelines to assist others in replicating our work. Large local and worldwide initiatives in neuroscience have lately arisen, such as the Swiss Blue Brain Project, the European Human Brain Project, the Allen Brain Observatory, and the American BRAIN initiative, which are pushing traditional neuroscience toward the big science paradigm.

These initiatives, at their core, use big data to simulate the functioning of the brain in great detail, down to the level of individual neuron types. Characterization of various elements, including as neuron types, synapses, and ion channels, is required for the data-driven method used for such large-scale modelling. Although the structural and functional characteristics of these entities may be assessed in the laboratory to some extent, the complexity and cross-scale nature of the simulated events preclude a thorough assessment of all factors at play. As a result, experimental results must be supplemented with global scientific knowledge recorded in the academic literature. The necessity for big, high-quality databases of literature-curated information regarding things linked to neuroscience is becoming increasingly apparent as a result of this scenario. We also need to be able to accurately catalogue previously published experimental observations so that they may be utilised to incorporate as values for modelling parameters or to validate emerging model features. Many efforts have been made in the field of neuroscience to categorise and characterise the many sorts of entities reported in the literature. A novel manual curation methodology has recently been developed to enhance traceability and reusability of systematically curated literature for the vast number of parameters required for thorough data-driven modelling of the brain.

***Correspondence to:**

Shawn Kruger
Editorial Office
Journal of Neuroinformatics and Neuroimaging
London
United Kingdom
E-mail: i_kruger51@gmail.com