Research Grants 2019- 2021

Title: Documenting word order variation in Mayan languages: A collection of Ch'ol narratives

Investigators: Jessica Coon

Duration: 2018 2019

Funding Agency: National Geographic Society, Explorers Grant

Summary:

Today there are thirty distinct Mayan languages spoken in Mexico, Guatemala, and Belize. As descendants of a common ancestor language, known as Proto-Mayan, these contemporary languages have for millennia maintained many of the distinctive grammatical characteristics of their common ancestor language, and have managed to do so in the face of centuries of contact with the dominant colonial language of the region, Spanish. However, with an increase in globalization and exposure to media of all forms, language diversity worldwide is diminishing at an alarming rate. Unless radical changes are made,

Title: **Cognitive predictors of second language learning success**

Investigator: **Meghan Clayards**

Duration: **2021-2026**

Funding Agency: **NSERC**

Summary:

Learning to distinguish between nonnative speech sounds can be challenging, especially when those sounds make distinctions that are not part of the native language or rely on particular acoustic-phonetic cues in a way that is not part of the native language. While there is generally learning over time, these

Title: **Compositional reasoning and OOD generalization in multimodal transformer models**

Investigators: **Yash Goyal, Aishwarya Agarwal, Siva Reddy, Aaron Carouville (PI)**

Durations: **2021 2023**

Agency: **Samsung-Mila Collaboration Grant**

Summary: Recent large scale transformer based language models have demonstrated that they can learn image representations from scratch using natural language supervision and can achieve remarkable zero-shot image-classification performance. However, in spite of being pretrained on large scale image-caption datasets and having millions of parameters, it is being revealed that such multimodal models lack essential capabilities that we would expect in an AI agent, such as understanding and reasoning about the compositional nature of language and vision data, and the ability to generalize to out-of-distribution (OOD) data.

This project proposes to systematically evaluate and improve the compositional reasoning and OOD generalization abilities of one of the flagship systems – CLIP (Computational Linguistics and Psycholinguistics)

Title: **Advanced computing infrastructure for integrating machine learning and**

Investigators:

Durations: **2021**

Agency:

Durations: **September 2020 August 2022**

Agency: **IVADO Fundamental Research Grant**

Summary: The main research objective in this project is to improve the state of the art for structured prediction using neural networks. The use of constraint programming during the training and inference phases of neural networks (NNs) will be investigated. For this project, Natural Language Processing (NLP) is considered as domain rich in structured prediction problems, and the focus is on parsing natural language to semantic structures such as semantic role labeling (the problem of parsing text to who did what to whom) and knowledge graph query prediction (the problem of parsing language to machine-executable representations) in order for the research project can have practical impact. In particular, the focus is on texts in Finance, Health and News domains for semantic role labeling, and Encyclopedic texts for knowledge graph queries.

Title: **Access to resources on COVID-19 through a chatbot**

Investigators: **Dialogue, Mila, Nu Echo, Samasource, Google, Johns Hopkins University and Data performers**

Durations: **May 2020 Sep 2020**

Agency: **Scale AI**

Summary: This project is to build, deploy and scale the chatbot infrastructure and interaction designers to create effective user experiences. The current chatbot's capacity is to be augmented by the expertise available at Nu Echo. Nu Echo has available capacity to significantly contribute to this project, and the track record of building production-ready conversational systems using technology in use by Dialogue (Rasa platform).

A large-scale language model (BERT) is pre-trained on COVID content and fine-tuned on question-answering datasets. The answers are encoded using the language model in dense vectors. The questions