**Deep Learning-Based Multi-Modal Indexing of Heterogeneous Clinical Data for Patient Cohort Retrieval**

Sanda M. Harabagiu, PhD1, Travis Goodwin1

1Human Language Technology Research Institute and Department of Computer Science, University of Texas at Dallas, U.S.A.

**SESSION TOPIC: BioCaddie and Resource Indexing**

Heterogeneity is one of the attributes of clinical Big Data, as the clinical picture of patients is documented by narratives, images, signals, etc. The ability to search through such data is made possible by multi-modal indexes, capturing the knowledge processed across all forms of clinical documents. Taking advantage of state-of-the-art deep learning methods, we have been able to generate a multi-modal index operating on a vast collection of electroencephalography (EEG) signal recordings and EEG reports and employ it is a patient cohort retrieval system. When EEG reports were indexed, the sections of the EEG reports were identified and medical language processing was performed. When the EEG signal recordings were processed, they were represented by EEG signal *fingerprints* as low-dimensional vectors produced by deep learning methods on the Big Data of EEG signals. Moreover, we organize the EEG fingerprints into a similarity-based hierarchy, which was included in the multi-modal index.

We used the multi-modal index in a patient cohort retrieval system, called MERCuRY (Multi-modal EncephalogRam patient Cohort discoveRY), which relies on medical language processing to identify the inclusion and exclusion criteria from the queries generated by neurologists. Using state-of-the-art relevance models adapted for a multi-modal index, we obtained very promising results, indicating that the multi-modal index is bridging the gaps between the electrode potentials recorded in the EEG signal and the clinical findings documented in the EEG reports.

Submitting Author’s Career Stage: Professor