Abstract

This paper presents an overview of and research utilizing the Brain Treebank dataset. Brain Treebank is a dataset of neurological, electroencephalographic (EEG) recordings taken from movie-watching test subjects, provided by the MIT Computer Science and AI Lab. The dataset has a sample size of 10 subjects, a cumulative recording time of 43 hours across 26 movies, and is currently the largest dataset of intracranial resolution with linguistic annotations (Wang et al., 2024). Thus, research was conducted to assess the usability of the novel Brain Treebank dataset by utilizing the dataset to conduct language decoding research. Natural language decoding is a process that is important for developing brain-computer interfaces and essentially consists of utilizing computational methods to decipher spoken or heard words and sentences from neurological recordings. The research presented in this paper sought to decode heard language from neurological recordings in the Brain Treebank dataset. Using the Python programming language, the dataset was analyzed for the presence of previously known linguistic neurological features, such as the N400. While the results were not significant enough to classify the produced models as accurate natural language decoders, the models were able to more reliably identify basic syntax and sentence structure (Wang et al., 2024). Therefore, these results demonstrate Brain Treebank's usability for neurolinguistic research and implies a requirement for a more precise dataset, a larger dataset, or better methods of analysis to create a true language decoding machine.