Reduction of Language Model Data Bases: Overcoming Chabot Implementation Obstacles

Nicholas J. Kaimakis
ICT/Univ. of Southern California
Los Angeles, California
kaimakis@usc.edu

Dan M. Davis
HPC-Education/USC
Long Beach, California
dmdavis@acm.org

ABSTRACT

This paper addresses optimization of language model data bases for use in Chatbots. The manifest utility and resultant proliferation of Chatbots is based on their users' ability to conversationally interface with computational assets. However, there is a concomitant pressure to reduce the size and increase the speed of client side devices. The currently widespread and rapidly growing acceptance of Chatbot approaches and the possibility of their implementation in DoD applications make this area vital to many who are engaged in Defense work. Not only is natural language processing a computationally expensive task, but storing language data takes up a sizeable amount of storage space and data base sizes do have an effect on search and retrieval times. A Chatbot must be able to understand and respond to a variety of unique user inputs and do so within the limits of conversationally tolerable latencies. The difficulty comes in trying to balance an effective chat agent while optimizing storage for peak performance. This paper discusses various methods of reducing language models, focusing on the Google News Word2vec model. Various methods were implemented in attempts to allow domain-specific analyzation and communication without the storage overhead of the significantly larger comprehensive models. We will document and analyze the history of improvements in similar efforts, set forth potential approaches, discuss our solutions within our environment, quantify impacts of the implementation of the selected approaches, outline future applications, and suggest topics for further research on the subject. The paper closes by offering the ways in which others can adopt this approach to their own efforts.

ABOUT THE AUTHORS

Nicholas Kaimakis is active in research at the Institute for Creative Technologies of the University of Southern California. His current research thrusts are in the use of computer generated avatars or video clips, animated and directed by natural language optimized Artificial Intelligence (A/I) programs that present a life-like dialogue capability to interact with remote users via the internet. He has demonstrable success designing efficient project structures, teaching programming, and managing multi-faceted teams. His current project is funded by the Navy and is designed to help improve knowledge of STEM fields across varied demographics with the development of an interactive interface that makes STEM information more accessible on-line. He is an undergraduate student studying Computer Science (CS) within the Viterbi School of Engineering at the University of Southern California.

Dan M. Davis is a consultant for the University of Southern California, focusing on large-scale distributed DoD simulations. Pre-retirement, for a decade he was the Director of USC's JESPP project for JFCOM. As the Assistant Director of the Center for Advanced Computing Research at Caltech, he managed Synthetic Forces Express, bringing HPC to DoD simulations. Prior experience includes serving as a Director at the Maui High Performance Computing Center and as a Software Engineer at the Jet Propulsion Laboratory and Martin Marietta. He has served as the Chairman of the Coalition of Academic Supercomputing Centers and has taught at the undergraduate and graduate levels. As early as 1971, Dan was writing programs in FORTRAN on one of Seymour Cray's CDC 6500's. He saw duty in Vietnam as a USMC Cryptologist and retired as a Commander, Cryptologic Specialty, U.S.N.R. He received B.A. and J.D. degrees from the University of Colorado in Boulder.