

Title: Statistical Method for Embedding-based Machine Translation

Abstract: This capstone project will explore the Multilingual Embedding-based Machine Translation using statistical methods. It considers two kindred Slavic Languages: Russian and Ukrainian and aims to learn such a linear transform  $W$  that minimizes the Euclidean distance between two embeddings for the respective word. It can be shown that a self-consistent linear mapping between semantic spaces should be orthogonal. Instead of making yet another regression problem, we can find optimal orthogonal transformation using singular value decomposition. It turns out that optimal transformation  $W'$  can be expressed via SVD components. The project's novelty lies in its departure from the prevalent deep learning paradigm, offering a fresh perspective on Multilingual Embedding-based Machine Translation through a rigorous statistical framework. Evaluation metrics, including precision at top-1, top-5, and top-10, will be employed to assess the quality and efficacy of the proposed approach.