

Application of Suffix Trees as an Implementation Technique for Varied-Length N-gram Language Models

Casey Redd Kennington

Suffix trees have been used in various applications in natural language processing because of their ability to represent sequential data in a way that is efficient to build, store, and access. How suffix trees are built and their properties are discussed as well as how this relates to language modeling. Language modeling is explained generally, and in some areas with more detail to illustrate how the suffix tree is applied to language modeling. The suffix tree language model is then compared to other existing language models in perplexity scores and memory usage, and further compared to one language model in several machine translation experiments. It is shown that the suffix tree language model is competitive with the state of the art in machine translation experiments and that suffix trees are well-suited for language modeling.