A Storage Reduction Method For Corpus-Based Language Models

Hsin-Hsi Chen and Yue-Shi Lee

Abstract

There are many progresses in corpus-based language models recently. However, the storage issue is still one of the major problems in practical applications. This is because the size of the training tables is in direct proportion to the parameters of the language models and the number of the parameters is in direct proportion to the power of these language models. In this paper, we will propose a storage reduction method to solve the problem that results from the large training tables. We use mathematical functions to simulate the distribution of the frequency value of the pairs in the training tables. For the good approximation, the pairs are grouping by their frequency. The experimental results show that although there is a little error rate introduced by the curve function, this scheme is still satisfactory because it performs the closed performance and no extra storage is required in pure curve-fitting model. Besides, we also propose a neural network approach to deal with the pairs classification which is a problem for all class-based approaches. The experimental results show that the neural network approach is suitable to deal with this problem in our storage reduction method.