Master's Thesis

Block coordinate descent methods for obtaining vector representations for words

Guidance

Professor Nobuo YAMASHITA

Ryota KATSUKI

Department of Applied Mathematics and Physics

Graduate School of Informatics

Kyoto University

February 2018

Abstract

In recent years, distributed representations of words have been widely utilized in the field of natural language processing. The idea of these representations is to assign a low-dimensional vector to each word, considering similarities and analogies with other words. In particular, the Global Vectors for word representation (GloVe) has attracted much attention as a model for getting high-performance distributed representations. GloVe obtains representations for words by solving a certain large-scale optimization problem whose variables are a bunch of vectors representing all words in huge documents. The optimization problem is solved by the stochastic gradient descent method. However, the method does not fully exploit the special structure of GloVe, which makes the convergence slow. Moreover, even if it finds some solution, the solution may not be accurate enough.

In this paper, we first propose a block coordinate descent method (BCD) that exploits the structure of GloVe's optimization problem. The objective function of the problem is squared sum of bilinear and linear functions. Thus it becomes a linear least squares problem when some variables are fixed. Since a solution of the linear least squares problem can be expressed explicitly, we can implement BCD efficiently. We show the global convergence of this method. Moreover, we perform numerical experiments, and show that the proposed method is faster than the existing ones. However, the performance of the distributed representations acquired by the proposed method is shown not to be very good. This is because the optimization problem in GloVe is a non-convex problem with a lot of local optima. The quality of distributed representations depends on the optimization method even if the optimal values are same.

In order to overcome this drawback, we also propose some improvements for the method after investigating the cause of this performance degradation. Moreover, assuming we know some word analogies in advance, we propose a new GloVe model that exploits such information. Finally, we give some numerical results, showing that distributed representations obtained by our improved method with the new GloVe model can achieve higher performance than the ones obtained by the existing approaches.