STATISTICAL-MECHANICS ANALYSIS OF DISTRIBUTED WORD EMBEDDINGS

HONGYIN LUO

Contents

1. Statistical-mechanics analysis of distributed word embeddings

We consider our corpora as a global system and different words are molecules in the system which move in stochastic way but effect each other by gravitation. Focus on two words which move freely: w and c. Assume the potential energy of w and c is $E_{w,c}$. Imagine the molecule of w is shot into the system. According to Boltzmann Distribution, the possibility of molecule w attracted by molecule c is

$$P_{system}(w,c) = \frac{e^{-\beta E_{w,c}}}{\sum_k e^{-\beta E_{w,k}}}$$

We also observe the words' distribution in our corpora. In our corpora, the possibility that the word w appear in the context window of word c is

$$P_{corpora}(w,c) = \frac{n(w,c)}{n(w)} \cdot \frac{|D|}{n(c)}$$

and the possibility that the word c appear in the context window of word w is

$$P_{corpora}(c,w) = \frac{n(c,w)}{n(c)} \cdot \frac{|D|}{n(w)} = P_{corpora}(w,c)$$

where $n(w,c)$ means the co-appearence time of w and c, $n(w)$ in the time of w's appearance, $n(c)$ is the time of c's appearance and $|D|$ is the length of the observed corpora.

Assume the total energy of the system is constant. then we can define a partition function which is constant valued function:

$$Q = \sum_k e^{-\beta E_{w,k}}$$

which indicate the total potential energy of word w and the system. Then the possibility of w attracted by c can be rewriten as

$$P_{system}(w,c) = \frac{1}{Q} \cdot e^{-\beta E_{w,c}}$$

Because the system is the corpora itself, so $P_{system}(w,c)$ and $P_{corpora}(w,c)$ should satisfy

$$P_{system}(w,c) = P_{corpora}(w,c)$$

then we obtain the following equation

$$E_{w,c} = -\frac{1}{\beta}(\log \frac{n(w,c) \cdot |D|}{n(w) \cdot n(c)} + \log Q)$$

Assume the frequency of w in the corpora is P(w), the frequency of c int the corpora is P(c) and the possibility that w and c appear in a context window is P(w,c). Then the PMI value of w and c is defined as

$$PMI_{w,c} = \log \frac{P(w,c)}{P(w) \cdot P(c)} = \log \frac{n(w,c) \cdot |D|}{n(w) \cdot n(c)}$$

Thus we can see PMI value of word w and c is proportional with the opposite number of the potential energy of the pair of words. As a result, the factorization of PMI matrix is a process to assign appropriate word vectors to fit the PMI or potential energy matrix. However, the common sense is the word vectors indicate the spatial location of a word. That's is correct because the gravitational potential energy can be calculated by distance. Assume m_w and m_c are the "mass" of word m and c.

$$E_{w,c} = -\frac{m_w \cdot m_c}{d} = -\frac{1}{\beta}(\log \frac{n(w,c) \cdot |D|}{n(w) \cdot n(c)} + \log Q)$$

d is the distance of word w and c. In most time the "mass" of words is ignored, in the other words, is simply set to 1 because it don't have apparent physical meanings. We found that $\frac{1}{d}$ is actually proportional with $PMI_{w,c}$,which indicates the smaller d is, the more similar is between two words.

Then we discuss the Gradient Descent algorithm applied in optimizing process. In fact, the gradient of potential is force. The update of word vectors is often like this:

$$w^{t+1} = w^t + \nabla f$$

where f is target function of the optimization problem. If we take above conclusions into consideration, the gradient descent algorithm adjusts the spatial location of a word under the effect of other words' gravitation until the system reaches an equilibrium. As a result of that, words "move linearly" while optimizing and the words' locations in the space are physically meaningful. The semantic information of a word is all contained the location of the word in semantic vector space.