TransE

- The approach in this paper is exactly the same as that of word2vec, and since KG embeddings + relations can naturally infer another vector, it is more intuitive compared to bag-of-words. It is essentially a direct application of word2vec's idea.
- Unlike word2vec, which builds a probabilistic model, this paper takes a more straightforward approach by modeling the embeddings as k-dimensional vectors (normalized to have unit norm), then optimizing the embedding vectors to minimize the loss.
- The loss function design draws on negative sampling, but in word2vec, a sigmoid function is used to map the inner product of vectors to the range of 0 to 1, thus reducing the impact of the norm on the loss. However, this paper does not employ such a mechanism and only **restricts the norm**? In the original paper, it was said that it was due to the fact that norm could have too much of an effect on distance. But, we know that in word embedding, the norm of embedding will also contain some information. And, how would the results change if we introduced a sigmoid function or trained the model with a norm-independent loss function?
- loss function in TranE

$$\sum_{(h,t),(h',t')\in T_{\text{batch}}} \nabla \left[\gamma + d(h+l,t) - d(h'+l,t')\right]_+$$

- The loss function in word2vec using negativing sampling is

$$Loss_{neg}(v_c, o, U) = -log(\sigma(\mu_o^T v_c)) - \sum_{k=1}^{K} log(\sigma(-\mu_k^T v_c))$$