**GloVe: Global Vectors for Word Representation**

**Jeffrey Pennington, Richard Socher, Christopher D. Manning**

Computer Science Department, Stanford University, Stanford, CA 94305

jpennin@stanford.edu, richard@socher.org, manning@stanford.edu

**Abstract**

Recent methods for learning vector space representations of words have succeeded in capturing fine-grained semantic and syntactic regularities using vector arith- metic, but the origin of these regularities has remained opaque. We analyze and make explicit the model properties needed for such regularities to emerge in word vectors. The result is a new global log- bilinear regression model that combines the advantages of the two major model families in the literature: global matrix factorization and local context window methods. Our model efficiently leverages statistical information by training only on the nonzero elements in a word-word co- occurrence matrix, rather than on the en- tire sparse matrix or on individual context windows in a large corpus. The model produces a vector space with meaningful sub-structure, as evidenced by its performance of 75% on a recent word analogy task. It also outperforms related models on similarity tasks and named entity recognition.

**摘要**

最近的一些学习词向量空间的表示方法成功地利用矢量算法来捕获细粒度语义和句法规律, 但这些规律的起源仍然很难理解。我们分析并明确了在词向量中出现这种规律所需的模型属性。分析结果得到一种新的全局对数双线性回归模型, 它结合了文献中两个主要模型体系的优点: 全局矩阵分解和局部上下文窗口方法。我们的模型只对word-word共现矩阵中的非零元素进行训练, 而不是在整个共现矩阵或一个巨大的语料库的单个上下文窗口，可有效地利用统计信息。该模型生成了一个带有含意义子结构的向量空间，在最近的一个词类比任务中表现为75%，还优于相似性任务和命名实体识别上的相关模型。