**GloVe: 全局词向量表示方法**

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

**计算机科学系，斯坦福大学, Stanford, CA 94305**

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

**摘要**

最近用于学习单词矢量空间表示的方法已经可以成功地使用矢量算术捕捉细粒度的语义和句法规则，但这些规律的得到过程仍然是晦涩难懂的。 我们分析并明确了这些在词向量中出现的规则所需的模型属性。 其结果是新的全局对数-双线性回归模型，它结合了文献中两个主要模型族的优点：全局矩阵分解方法和局部上下文窗口方法。 我们的模型通过仅对单词共生矩阵中的非零元素进行训练而非整个稀疏矩阵或大型语料库中的单个上下文窗口来有效利用统计信息。 该模型产生的有意义的子结构的向量空间，其最近的Word Analogy Task(Tomas Mikolov提出的测试集)上的性能为75％。 它在相似性任务和命名实体识别方面也优于相关模型。

**1引言**

一种语言的语义向量空间模型用实际值的向量来表示每个词。这些向量可以在各种应用中作为特征，如信息检索（Manning 等, 2008），文档分类（Sebastiani, 2002），问题回答（Tellex 等, 2003），命名实体识别（Turian等, 2010）和语法分析（Socher等, 2013）。大多数词向量方法依赖于一对词向量之间的距离或角度作为评估其单词表示的内在质量的主要方法。最近，Mikolov等人(2013c)引入了一种基于单词类比的新评估方案，该方案通过检查单词各个维度的差异而不是单词向量之间的标量距离来探究单词向量空间的更好的结构。例如，应该用向量方程 王 - 女王=男人 - 女人 在向量空间中编码类比“王之于女王如同男人之于女人”。这种评估方案倾向于生成有意义的维度的模型，从而捕捉分散表示的多种主要思想（Bengio, 2009）。用于学习单词向量的两个主要模型族为：1）全局矩阵分解方法，如潜在语义分析（LSA）（Deerwester等, 1990）; 2）局部上下文窗口方法，如Mikolov等人（2013c）的skip-gram模型。目前，这两个模型族都有严重的缺陷。虽然像LSA这样的方法有效地利用统计信息，但他们对单词类比任务的表现相对较差，表现出次优的向量空间结构。像skip-gram这样的方法在类比任务上可能会做得更好，但由于它们在单独的局部上下文窗口上进行训练，而不是在全局共现计数上，所以很难利用语料库的统计数据。在这项工作中，我们分析了产生线性意义维度所需的模型属性，并认为全局对数-双线性回归模型适合于这样做。我们提出了一个特定的加权最小二乘模型，训练全局单词共现计数，从而有效利用统计数据。该模型能够产生有意义的子结构的向量空间，正如其Word Analogy数据集的75％准确度的最新性能所证明的。我们还证明，我们的方法在几个词相似度任务上胜过其他现有方法，并且也取决于共同的命名实体识别（NER）基准。我们提供了模型的源代码以及预训练词向量在http://nlp. stanford.edu/projects/glove/.

**2相关工作**

**矩阵分解方法。**用矩阵分解方法生成低维词表示可以追溯到LSA。这些方法利用低秩近似分解大矩阵从而获取关于语料库的统计信息。由这些矩阵捕获的特定类型的信息因应用而异。在LSA中，矩阵是“词语-文档”类型的，即，行对应于单词，并且列对应于语料库中的不同文档。相比之下，例如Hyperspace Analogue to Language（HAL）（Lund和Burgess, 1996）利用了“词语-词语”类型的矩阵，即行和列对应于单词，并且元素对应单词出现在另一个单词的上下文中的次数。 HAL和相关方法的一个主要问题是，最频繁的词对相似性度量有不成比例的影响：例如，两个词与or , and共同出现的次数，将会对它们的相似性产生很大的影响，尽管它们的语义相关性相对较小。存在多种方法可以解决HAL的缺点，例如COALS方法（Rohde等, 2006），其中共现矩阵首先通过基于熵或基于相关性的归一化转换。这种类型转换的一个优点是，尺寸合理的可能跨越8或9个数量级的语料库的原始同现计数会被压缩，以便在更小的间隔内更均匀地分布。各种较新的模型也采用这种方法，包括一项研究（Bullinaria 和 Levy, 2007），该研究表明正点互信息（PPMI）是一种很好的转变。而最近，Hellinger PCA（HPCA）形式的平方根变换（Lebret和Collobert, 2014）被认为是学习单词表示的有效方式。