**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](mailto:manning@stanford.edu)

**译者 康宇辰**

**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。

**浅层的基于窗口方法。**另一种方法是学习可帮助在本地环境窗口内进行预测的单词表示。例如，Bengio等人（2003）引入了一个模型来学习单词向量表示，并作为语言建模的简单神经网络体系结构的一部分。 Collobert和Weston（2008）将矢量训练与下游训练目标分开，这为Collobert等人（2011）铺平了道路：使用单词的全部上下文来学习单词表示，而不像其他语言模型一样仅仅是前文。最近，完整神经网络结构对于学习有用的单词表示的重要性已经受到质疑。 Mikolov等人（2013a）的skip-gram和bag-of-words (CBOW)模型基于两个单词向量之间的内积提出了一个简单的单层体系结构。 Mnih和Kavukcuoglu（2013）也提出了密切相关的向量对数双线性模型，vLBL和ivLBL，并且Levy等人（2014）基于PPMI度量提出了明确的单词嵌入模型。在skip-gram和ivLBL模型中，目标是给定单词本身预测单词上下文，而CBOW和vLBL模型的目标是根据上下文预测单词。通过对词类比任务的评估，这些模型展示了将语言样本学习转换为词向量之间的线性关系的能力。与矩阵分解方法不同，浅层的基于窗口的方法的缺点是它不直接对语料库的共现统计进行操作。相反，它会扫描整个语料库中的上下文窗口，而无法利用其中的大量重复数据。

**3 GloVe模型**

语料库中单词出现的统计信息是所有用于学习词汇表示的无监督方法的主要信息来源，尽管现在存在许多这样的方法，但问题仍然是如何从这些统计生成意义，以及得到的单词向量可能代表什么意思。在这一节中，我们将对这个问题进行一些阐述。我们利用我们的见解，建立了一个新的用于单词表示的全局向量模型GloVe，因为全局语料库统计信息直接由模型捕获。首先我们创建一些符号。设单词共现计数矩阵用X表示，其中元素表示单词j出现在单词i的上下文中的次数。令 = 表示任何单词在单词i的上下文中出现的次数。最后，设= P(j |i)= / 为单词j出现在中的概率。

表1：目标单词ice和steam在从60亿标记语料库中挑选的上下文中的共现概率。只有在这样的比例下，来自像water和fashion这样的无法判断的词汇的干扰才会消失，从而使得大的值（远大于1）与单词ice特有的相关性更好，小的值（远小于1）与单词steam特有的相关性更好。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Probability and Ratio | k = solid | k = gas | k = water | k = fashion |
| P(k|ice) | 1.9× | 6.6 × | 3.0 × | 1.7 × |
| P(k|steam) | 2.2 × | 7.8 × | 2.2 × | 1.8 × |
| P(k|ice)/P(k|steam) | 8.9 | 8.5 × | 1.36 | 0.96 |

我们从一个简单的例子开始，展示了如何从共现概率中直接提取某些方面的含义。考虑两个字i和j表现在特别的兴趣方面；具体而言，假设我们对热力学方面的概念感兴趣，而我们可能需要i=ice和j=steam。这些词的关系可以通过研究它们与各种检测单词k的同现概率的比率来检验。K与ice有关而与steam无关，设k=solid，我们预计比率 / 将会很大。同样，对于k与steam有关但与ice无关，比如k＝gas，这个比例应该是小的。对于像水或时尚这样的词语，要么与ice和steam都有关，要么两者都不相关，这个比例应该接近于1。表1显示了这些概率及其对大型语料库的比率，数字证实了这些预期。与原始比值相比，该比值能够更好地区分无关词（ice和fashion）中的相关单词（solid和gas），并且还能更好地区分这两个相关单词。

上述观点表明，词向量学习的适当出发点应该是同现概率的比率，而不是概率本身。 注意到比率 / 取决于三个单词i，j和k，最一般的模型采取的形式是，

= / (1)

其中w∈是词向量，∈是独立的上下文词向量，其作用将在4.2节中讨论。在这个等式中，右侧是从语料库中提取到的，并且F可以依赖于一些尚未指定的参数。F的可能性数量很大，但通过设定一些必要条件，我们可以得到一个唯一的选择。首先，我们希望F在单词向量空间中对表示比率 / 的信息进行编码。由于向量空间本质上是线性结构，所以最自然的方法是使用向量差。有了这个目标，我们可以将我们的目光放在在那些仅依赖于两个目标词的差异的函数F上，修改方程(1)至，

= / (2)

接下来，我们注意到等式(2)中F的参数是向量，而右边是标量。虽然F可以被看作是由例如神经网络参数化的复杂功能，但这样做会混淆我们试图捕获的线性结构。为了避免这个问题，我们可以首先拿出参数的点积，

= / (3)

这防止F以不希望的方式混合向量维度。接下来，请注意，对于单词共现矩阵，单词和上下文单词之间的特征是随机的，我们可以自由地交换这两个角色。为了这样做，我们不仅要交换，还要交换。 我们的最终模型在这种情况下应该是不变的，但是等式(3)不是。 但是，对称性可以分两步恢复。 首先，我们要求F是组（，+）和（，×）之间的同态，即，

= / (4)

又根据等式(3)有，

= = / (5)

等式(4)的解是F = exp，或者，

= = (6)

接下来，我们注意到等式(6)将表现出交换对称性，如果没有右边的。 然而，这个项与k无关，所以它可以合并到的偏差中。最后，为增加一个额外的偏差量恢复对称性，

+ = (7)

等式(7)对等式(1)进行了极大的简化，但它实际上是不明确的，因为只要其参数为零，对数就会发散。 解决这个问题的一个办法是在对数中包含一个加法移位，→，它保持X的稀疏度，同时避免发散。将共生矩阵的对数分解的想法与LSA密切相关，我们将在实验中使用所得模型作为基准。这种模式的一个主要缺点是，它平等地衡量所有共现情况，即使是那些很少发生或从未发生过的。这种罕见共现是具有干扰的，比常见的信息携带更少的信息-然而零条目甚至可以占X数据的75-95％，这取决于词汇量规模和语料库。

我们提出了一个新的加权最小二乘回归模型来解决这些问题。将等式(7)作为最小二乘问题，并将加权函数引入到代价函数中得出我们的模型

J = (8)

其中是词汇量的规模大小。加权函数应该遵守以下属性：

1. = 0。如果被看作一个连续函数，它应该消失，因为的速度足够快，而是有限的。

2. 应该是非递减的，这样罕见共现不会过多。

3. 对于大的值，应该相对较小，这样频繁共现不会过多。

当然，大量函数满足这些性质，但是我们发现可以很好地工作的一类函数可以参数化为，

(9)

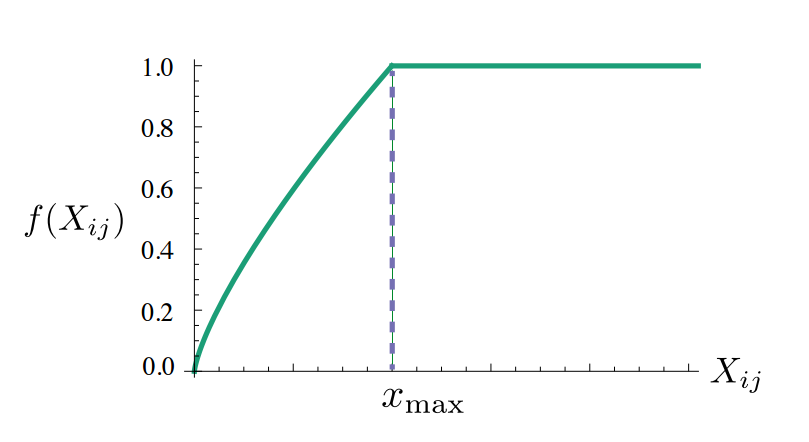


图1：加权函数f，α= 3/4

模型的性能取决于截断点，我们在所有实验中都将 = 100固定。我们发现α= 3/4在α= 1的线性版本上略有改进。尽管我们只提供了选择3/4值的经验动机，但有趣的是，发现类似的分数比例缩放可以得到最好的性能 （Mikolov等，2013a）。

**3.1与其他模型的关系**

由于所有用于学习单词向量的无监督方法最终都基于单词在语料库中的出现统计，所以各种模型之间应该有共同点。尽管如此，某些模型在这方面仍然有些不透明，特别是最近基于窗口的方法，如skip-gram和ivLBL。因此，在本小节中，我们将展示这些模型如何与我们提出的模型相关，如等式（8）。

skip-gram或ivLBL方法的出发点是一个模型，表示单词j出现在单词i的上下文中的概率。具体来说，让我们假设是softmax函数，

= / (10)

这些模型的大部分细节与我们的目的无关，除了他们试图在上下文窗口扫描整个语料库时最大化对数概率这一事实。训练以一个在线随机的方式进行，但隐含的全局目标函数可以写成：

J = - (11)

为这个总和中的每个项评估softmax归一化因子是昂贵的。为了进行有效的训练，skip-gram和ivLBL模型引入了的近似值。但是，如果我们首先将那些具有相同i和j值的项组合在一起，则可以更高效地计算等式(11)的和，

J = - (12)

当我们使用了由共生矩阵X给出的近似项的数量。

回想一下 = 和 = / 的符号，我们可以将J重写为，

J = - = (13)

其中是我们与类比定义的分布点和的交叉熵。作为交叉熵误差的加权和，这个目标与等式(8)的加权最小二乘目标有一些形式上的相似之处。 事实上，可以直接优化等式(13),而不是在skip-gram和ivLBL模型中使用的在线训练方法。我们可以将这个目标解释为“global skip-gram”模型，进一步研究下去可能会很有意思。另一方面，等式(13)展示了一些不合需要的特性，在将其作为学习单词向量的模型之前应该加以解决。

首先，交叉熵误差只是概率分布之间的许多可能的距离度量之一，它有一个令人遗憾的特性，即长尾分布往往建模不良，对不太可能的事件给予过多的权重。 此外，为了边界有限，它要求模型分布Q被正确地归一化。这提出了计算瓶颈，由于等式(10)中的整个词汇的总和，可能要考虑不需要Q的这种性质的一种的不同距离度量。自然的选择是最小二乘任务，其中Q和P中的归一化因子被丢弃，

= (14)

其中= 和 = 是非标准化的分布。在这个阶段会出现另一个问题，即常常需要非常大的值，这可能会使优化复杂化。有效的修改办法是尽量减少和的对数的平方误差，

= = (15)

最后，我们观察到尽管加权因子是通过skip-gram和ivLBL模型固有的在线训练方法预先确定的，但它并不保证是最优的。事实上，Mikolov等人（2013a）观察到，通过过滤数据可以提高性能，从而降低频繁词的加权因子的有效值。考虑到这一点，我们引入了一个更一般的权重函数，让我们可以自由地依赖上下文单词。结果是，

= (16)

这相当于等式(8)的成本函数[[1]](#footnote-1)，我们先前得到了这个函数。

**3.2模型的复杂性**

从等式(8)和加权函数的形式可以看出，模型的计算复杂度取决于矩阵X中非零元素的数量。由于这个数字总是小于矩阵的条目总数，模型的规模不会比O差。乍一看，这可能看起来像是基于浅层窗口方法的实质性改进，这种方法随着语料库大小进行缩放。但是，典型的词汇有数十万个单词，因此可能在数千亿，实际上比大多数语料库要大得多。出于这个原因，确定是否对X的非零元素的数目上设置更严格的界限是很重要的。

为了对X中的非零元素的数量做出具体的陈述，我们有必要对字词共现的分布做出一些假设。具体而言，我们假定可以将单词i与单词j的共现出现的次数模拟为该对单词的频率等级的幂函数:

= k / (17)

语料库中的词的总数与共生矩阵X的所有元素的和成比例，

∼ = = k (18)

当我们重写最后的和为广义谐波数。总和的上限是最大频率等级，其与矩阵X中的非零元素的数目一致。该数目也等于等式(17)中的r的最大值，使得≥1，即 = 。因此我们可以将等式(18)写成，

∼ (19) 我们对与如何相关感兴趣，当两个数字都很大时；因此我们可以自由地扩大等式右边来增大。为此我们使用广义谐波数的扩展（Apostol，1976），

= /1-s +ζ + O if s > 0, s ≠ 1 (20)

得到，

∼ +ζ + O (21)

对于本文所研究的语料库，我们观察到可以通过等式(17)得到很好的模拟当α= 1.25。在这种情况下，我们有 = O。因此，我们得出结论：模型的复杂性比最差情况O好得多，实际上它比基于窗口的在线方法比如O要好一些。

**4实验**(该部分未译)

**5结论**

最近，相当多的关注集中在分布式的单词表示最好从基于计数的方法还是基于预测的方法中学习的问题上。目前，基于预测的模型获得了大量支持;例如，Baroni等人（2014）认为，这些模型在一系列任务中表现更好。在这项工作中，我们认为这两类方法在基本水平上并没有显着差异，因为它们都探测了语料库的潜在共现统计数据，但基于计数的方法捕获全局统计数据的效率可能是更好的。我们构建了一个利用计数数据这一主要优点的模型，同时捕获了最新的基于预测的对数双线性方法如word2vec中常见的有意义的线性子结构。其结果，GloVe，是一个新的全局对数双线性回归模型，用于无监督地进行单词表示的学习，该模型在词类比，单词相似性和命名实体识别任务方面优于其他模型。

**致谢、参考文献**未译

1. 我们也可以包含等式(16)中的偏差项。 [↑](#footnote-ref-1)