GloVe：用于单词表示的全局向量

Jeffrey Pennington, Richard Socher, Christopher D. Manning
斯坦福大学计算机科学系，斯坦福，加利福尼亚州 94305
jpennin@stanford.edu, richard@socher.org, manning@stanford.edu

摘要

最近的单词向量空间表征学习方法成功地利用向量度量捕捉到了细粒度的语义和句法规律性，但这些规律性的来源仍然不清楚。我们分析并阐明了词向量中出现这种规律性所需的模型属性。其结果是建立了一个新的全局对数双线性回归模型，该模型结合了文献中两个主要模型系列的优点：全局矩阵因式分解和局部语境窗口方法。我们的模型只对词-词共现矩阵中的非零元素进行训练，而不是对稀疏矩阵或大型语料库中的单个上下文窗口进行训练，从而有效地利用了统计信息。该模型生成的向量空间具有有意义的子结构，在最近的一项单词类比任务中，该模型的表现达到了 75%。该模型在类比任务和命名实体识别上的表现也优于相关模型。

1 简介

语言的语义向量空间模型用实值向量表示每个单词。这些向量可用作各种应用中的特征，如信息检索（Manning 等人，2008 年）、文档分类（Sebastiani，2002 年）、问题解答（Tellex 等人，2003 年）、命名实体识别（Turian 等人，2010 年）和解析（Socher 等人，2013 年）。

大多数词向量方法都依赖于词向量对之间的距离或角度，以此作为评估词表示集内在质量的主要方法。最近，Mikolov 等人（2013c）提出了一种基于词类的新评估方案，该方案可探测通过不考虑词向量之间的标量距离，而是考虑它们之间不同维度的差异，来研究词向量空间的精细结构。

例如，"国王之于王后，就像男人之于女人 "这一类比应在向量空间中通过向量方程*国王 − 王后 = 男人 − 女人*。这种评估方案有利于产生均值维度的模型，从而捕捉到分布式表征的多聚类思想（Bengio，2009 年）。

学习词向量的两个主要模型系列是1) 全局矩阵因式分解方法，如潜在语义分析（LSA）（Deerwester 等人，1990 年）；2) 局部语境窗口方法，如 Mikolov 等人（2013 年 c）的跳格模型。目前，这两类方法都存在显著缺陷。虽然 LSA 等方法能有效地利用统计信息，但它们在单词类比任务中的表现却相对较差，说明向量空间结构不够理想。跳格（skip-gram）等方法可能在类比任务中表现较好，但由于它们是根据单独的局部上下文胜点而非全局共现计数进行训练，因此对相关词的统计信息利用不充分。

在这项工作中，我们分析了产生线性意义方向所需的模型属性，并认为全局对数-线性回归模型适合于产生线性意义方向。我们提出了一种特殊的加权最小二乘法模型，该模型以全局词-词共现计数为基础进行训练，从而有效地利用了统计数据。该模型生成的词向量空间具有有意义的子结构，其在词类比数据集上 75% 的最高准确率就是最好的证明。我们还证明，我们的方法在多项词语相似性任务以及常见的命名词识别（NER）基准任务中的表现优于其他现有方法。

我们在 http://nlp.stanford.edu/projects/glove/ 网站上提供了模型的源代码和训练过的词向量。

2 相关工作

矩阵因式分解方法。用于生成低维词表示的矩阵因式分解方法的起源可以追溯到 LSA。这些方法利用低秩近似法来分解大型矩阵，从而获取语料库的统计信息。这些矩阵所捕捉的特定信息类型因应用而异。在 LSA 中，矩阵属于 "术语-文档 "类型，即行对应单词或术语，列对应语料库中的不同文档。与此相反，超空间类比语言（HAL）（Lund 和 Burgess，1996 年）等则使用 "词-词 "类型的矩阵，即行和列与词相对应，条目与给定词在另一个给定词的上下文中出现的次数相对应。

HAL 和相关方法的一个主要问题是，出现频率最高的词对相似性测量的贡献不成比例：例如，两个词与 或 *和 的*共现次数会对它们的相似性产生很大影响，尽管对它们的语义相关性的影响相对较小。有许多技术可以解决 HAL 的这一缺陷，例如 COALS 方法（Rohde 等人，2006 年），该方法首先通过基于熵或相关性的归一化对共现矩阵进行转换。这种转换的优点在于，原始的共现计数（对于一个合理大小的语料库来说，共现计数可能会跨越 8 或 9 个数量级）会被压缩，从而在一个较小的区间内分布得更加均匀。各种较新的模型也采用了这种方法，其中一项研究（Bullinaria 和 Levy，2007 年）指出，正向点偶信息（PPMI）是一种很好的变换方法。最近，海灵格 PCA（HPCA）形式的平方根式转换（Lebret 和 Collobert，2014 年）被认为是学习单词表征的有效方法。

基于浅窗口的方法。另一种方法是学习单词表征，帮助在局部语境窗口内进行预测。例如，Bengio 等人（2003 年）推出了一个学习词向量表征的模型，作为语言建模的简单神经网络架构的一部分。Collobert 和 Weston（2008 年）将词向量训练与下游训练目标解耦，为 Collobert 等人的研究铺平了道路。这为 Collobert 等人（2011 年）使用单词的全部上下文来学习单词代表，而不是像语言模型那样只使用前面的上下文铺平了道路。

最近，完整神经网络结构对于学习有用单词代表的重要性受到了质疑。Mikolov 等人（2013a）的跳格和连续词袋（CBOW）模型提出了一种基于两个词向量内积的简单单层结构。Mnih 和 Kavukcuoglu（2013 年）也提出了密切相关的向量对数线性模型 vLBL 和 ivLBL，Levy 等人（2014 年）提出了基于 PPMI 指标的显式词嵌入模型。

在 skip-gram 和 ivLBL 模型中，目标是根据单词本身预测单词的上下文，而 CBOW 和 vLBL 模型的目标则是根据单词的上下文预测单词。通过对单词类比任务的评估，这些模型展示了学习单词向量间线性关系的语言模式的能力。

与矩阵因式分解方法不同，基于浅层窗口的方法的缺点是不能直接利用语料库中的共现统计数据。相反，这些模型在整个语料库中扫描上下文窗口，无法利用数据中的大量重复。

3 GloVe 模型

语料库中单词出现的统计数据是所有无监督学习单词代表方法的主要信息来源，尽管现在有许多这样的方法，但问题仍然是如何从这些统计数据中生成意义，以及由此产生的单词向量如何代表意义。在本节中，我们将对这一问题作一些说明。我们利用自己的洞察力构建了一个新的词语表征模型，我们称之为 GloVe（全局向量），因为该模型直接捕捉到了语料库的全局统计数据。

首先，我们建立一些符号。让词-词共现计数矩阵用 X 表示，其条目 X_{ij} 表示词 j 在词 i *的*上下文中出现的次数。让 $X_i = \sum_k X_{ik}$ 表示任何词在词 i 的上下文中出现的次数。最后，让 $P_{ij} = P(j \mid i) = X_{ij} / X_i$ 表示词 j 在词 i 的上下文中出现的概率。

表 1：目标词 "*冰* "和 "*蒸汽* "与 60 亿标记语料库中选定语境词的共现概率。只有在比值中，来自*水*和*时尚*等非区分词的噪音才会被抵消，因此，大值（远大于 1）与冰的特定属性相关，小值（远小于 1）与蒸汽的特定属性相关。

概率和比率	k= *固体*	k= *气体*	k= *水*	K= *时尚*
$P(k\|$ *冰*$)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|$ *蒸汽*$)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|$ *冰*$)/P(k\|$ *蒸汽*$)$	8.9	8.5×10^{-2}	1.36	0.96

单词 i 的上下文

我们先举一个简单的例子，说明如何直接从共现概率中提取意义的某些方面。具体来说，假设我们对 "热力学相位 "这一概念感兴趣，我们可以将 i= *ice* 和 j= *steam* 视为热力学相位。我们可以通过研究这些词与不同探究词 k *的*共现概率之比来研究它们之间的关系。对于与冰相关但与蒸汽无关的词 k，例如 k= *solid*，我们预计 P_{ik}/P_{jk}的比率会很大。同样，对于与蒸汽有关但与冰无关的词 k，如 k= *gas*，比值应该很小。对于像*水*或*时尚*这样既与冰又与蒸汽相关或两者都不相关的词 k，比率应该接近于 1。表 1 显示了大型语料库中的这些概率及其比率，这些数字证实了上述预期。与原始概率相比，比率更能区分相关词（*固体*和*气体*）和不相关词（*水*和*时尚*），也更能区分两个相关词。

上述论证表明，词向量学习的适当起点应该是共同出现概率的比率，而不是概率本身。由于比值 P_{ik}/P_{jk}取决于三个词 i、j 和 k，因此最一般的模型形式为：

$$F(w_i, w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}, \qquad (1)$$

其中 $w\in \mathrm{R}^d$ 是词向量，$\tilde{w}\in \mathrm{R}^d$ 是独立的上下文词向量，其作用将在第 4.2 节中讨论。在这个等式中，右边是从语料库中提取的，而 F 可能取决于一些尚未指定的参数。F 的可能性非常多，但通过强制执行一些必要条件，我们可以选出一个唯一的选择。首先，我们希望 F 编码信息呈现词向量空间中的比率 P_{ik}/P_{jk}。由于向量空间本质上是线性结构，因此最自然的方法就是向量差分。为此，我们可以将考虑范围限制在仅取决于两个目标词差值的函数 F 上，将公式 (1) 修改为：

$$F(w_i - w_j, \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}. \qquad (2)$$

接下来，我们注意到公式 (2) 中 F 的参数是向量，而右侧是标量。虽然可以将 F 视为由神经网络等参数化的复杂函数，但这样做会模糊我们试图捕捉的线性结构。为了避免这个问题，我们可以先取参数的点积、

$$F((w_i - w_j)^T \tilde{w}_k) = \frac{P_{ik}}{P_{jk}}, \qquad (3)$$

这可以防止 F 以不可取的方式混合向量维数。接下来，我们要注意的是，对于词-词共现矩阵来说，词和上下文词的区分是任意的，我们可以自由地交换这两个角色。要做到这一点，我们不仅要交换 $w\ \tilde{w}$，还要交换 $X \to X^T$。我们的最终模型在这种重新标注下应该是不变量，但公式 (3) 却不是。不过，对称性可以分两步恢复。首先，我们要求 F 是同态的 $(\mathrm{R}, +)$ 组和 $(\mathrm{R}_{>0}, \times)$ 组之间，即

$$F((w_i - w_j)^T \tilde{w}_k) = \frac{F(w_i^T \tilde{w}_k)}{F(w_j^T \tilde{w}_k)}, \qquad (4)$$

根据公式 (3)，它的解法是

$$F(w_i^T \tilde{w}_k) = P_{ik} = \frac{X_{ik}}{X_i}. \qquad (5)$$

公式 (4) 的解是 F= exp，或、

$$w_i^T \tilde{w}_k = \log(P_{ik}) = \log(X_{ik}) - \log(X_i)\ 。 \qquad (6)$$

接下来，我们注意到，如果没有右侧的 $\log(X_i)$，公式 (6) 将表现出变化对称性。不过，这个项与 k 无关，因此可以将其吸收为 w_i的偏置 $b_{(i)}$。最后，为 $\tilde{w}_k$添加一个额外的偏置 $\tilde{b}_k$就恢复了对称性、

$$w_i^T \tilde{w}_k + b_i + \tilde{b}_k = \log(X_{ik}) \,. \tag{7}$$

式（7）是对式（1）的极大简化，但实际上定义不清，因为只要其参数为零，对数就会发生变化。解决这个问题的一个办法是在对数中加入加法移动，即 $\log(X_{ik})$ $\log(1+ X_{(ik)})$ ，这样既能保持 X 的稀疏性，又能避免对数畸变。将共生矩阵的对数因子化的想法与 LSA 密切相关，我们将在实验中使用由此产生的模型作为基线。该模型的一个主要缺点是它会对所有共现现象进行同等权重，即使是那些很少发生或从未发生的共现现象也不例外。这种罕见的共现是有噪声的，与频繁出现的共现相比，所携带的信息更少--然而，根据词汇量和语料库的不同，即使是零条目也占了 X 中数据的 75-95%。

我们提出了一种新的加权最小二乘回归模型来解决这些问题。将式 (7) 看作最小二乘法问题，并在成本函数中引入加权函数 $f(X_{(i)(j)})$，就得到了以下模型

$$J = \sum_{i,j=1}^{V} f(X_{(i)(j)}) \left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{(i)(j)} \right)^2, \tag{8}$$

其中 V 是词汇量的大小。权重

权重函数应符合以下特性：

1. $f(0) = 0$。如果把 f 看作连续函数，那么当 $x \to 0$ 时，它应该消失得足够快，以至于 $\lim_{x \to 0} f(x) \log^2 x$ 是有限的。
2. $f(x)$ 应该是不递减的，这样稀有的共现现象就不会被加权。
3. $f(x)$ 应该相对较小。
   这样，频繁出现的共现现象就不会被加权。

当然，有很多函数都能满足这些特性，但我们发现有一类函数效果很好，可以将其参数化为

$$f(x) = \begin{cases} (x/x_{\max})^{\alpha} & \text{如果 } x < x_{\max} \\ 1 & \text{否则} \end{cases} \tag{9}$$

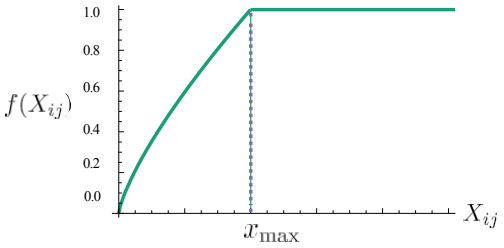


图 1：加权函数 f，α= 3/4。

模型的性能微弱地依赖于截止值，我们在所有实验中都将截止值固定为 $x_{\max} = 100$。我们发现 α= 3/4 比 α= 1 的线性版本有适度改善。虽然我们只是根据经验选择了 3/4 的值，但有趣的是，在 Mikolov 等人的研究（Mikolov et al.）

3.1 与其他模型的关系

由于所有用于学习词向量的无监督方法最终都是基于语料库的词出现率统计，因此模型之间应该存在共性。然而，某些模型在这方面仍然有些不透明，尤其是最近出现的基于窗口的方法，如 skip-gram 和 ivLBL。因此，在本小节中，我们将展示这些模型与我们提出的模型（如公式 (8) 所定义）之间的关系。

skip-gram 或 ivLBL 方法的出发点是以下概率的模型 $Q_{(i)(j)}$

为了简洁起见，我们假设 $Q_{(i)(j)}$是软最大值、

$$Q_{ij} = \frac{\exp(w_i^T \tilde{w}_j)}{\sum_{k=1}^{V} \exp(w_i^T \tilde{w}_k)}. \tag{10}$$

这些模型的大部分细节都与我们的目的无关，除了它们试图在文本窗口扫描语料库时最大化对数概率这一事实之外。训练是以在线随机方式进行的，但全局目标函数可以写成

$$J = -\sum_{\substack{i \in \text{语料库} \\ j \in \text{context}(i)}} \log Q_{(i)(j)}. \tag{11}$$

为这一总和中的每个词评估软最大值的归一化因子成本很高。为了降低成本以提高训练效率，跳格和 ivLBL

模型引入了 $Q_{(i)(j)}$的近似值。然而，如果我们先将 i 和 j 值相同的项归并到一起，那么公式 (11) 中的总和就能更有效地计算出来。

不过，如果我们先将 i 和 j 值相同的项归为一组，公式 (11) 中的总和就能更有效地求得、

$$J = -\sum_{i=1}^{V}\sum_{j=1}^{V} X_{ij}\log Q_{(i)(j)}、 \tag{12}$$

其中，我们利用了同类项的数量由共生矩阵 X 给出这一事实。回顾我们对 $X_i = \sum_k X_{ik}$ 和 $P_{(i)(j)} = X_{(i)(j)}/X_{(i)}$，我们可以将 J 改写为、

$$J = -\sum_{i=1}^{V} X_i \sum_{j=1}^{V} P_{(i)(j)}\log Q_{(i)(j)} = \sum_{i=1}^{V} X_i H(P_i, Q_{(i)}), \tag{13}$$

其中，$H(P_i, Q_i)$ 是词库的交叉熵。

我们通过类比 X_i 来定义分布 P_i 和 $Q_{(i)}$。作为交叉熵误差的加权和，该目标与公式 (8) 中的加权最小二乘法目标在形式上有些相似。事实上，可以直接优化公式 (13)，而不是采用跳格和 ivLBL 模型中的在线训练方法。我们可以将这一目标视为 "全局跳格 "模型，并对其进行进一步研究。另一方面，公式（13）显示了一些不理想的特性，在将其作为学习词向量的模型之前，应该先解决这些问题。

首先，交叉熵误差只是概率分布之间众多可能的距离度量中的一种，而且它有一个不幸的特性，即具有长尾的分布建模不佳，对不可能事件的权重过大。此外，要使该指标有界，需要对模型分布 Q 进行适当的归一化处理。在公式 (10) 中，由于需要对整个词汇进行求和，这就造成了计算上的瓶颈，因此最好考虑一种不同的距离测量方法，而不需要 Q 的这一特性、

$$J' = \sum_{i,j} X_{(i)}\left(\hat{P}_{(i)(j)} - \hat{Q}_{ij}\right)^2 \tag{14}$$

其中，$\hat{P}_{i(j)} = X_{(i)(j)}$ 和 $\hat{Q}_{ij} = \exp(w^T\tilde{w}_{(j)})_i$ 是非正态分布。在此阶段，另一个出现的问题是，$X_{(i)(j)}$ 的值往往非常大，这会使优化过程变得复杂。值非常大，这会使优化工作复杂化。

有效的补救措施是而不是 $\hat{P}$ 和 $\hat{Q}$ 的对数的平方误差、

$$\hat{J} = \sum_{i,j} X_i\left(\log\hat{P}_{ij} - \log\hat{Q}_{ij}\right)^2 = \sum_{i,j} X_{(i)}\left(w_i^T\tilde{w}_j - \log X_{(i)(j)}\right)^2. \tag{15}$$

最后，我们注意到，虽然加权因子 X_i 是由 skip-gram 和 ivLBL 模型固有的在线训练方法预先确定的，但它绝不保证是最优的。事实上，Mikolov 等人（2013a）发现，可以通过过滤数据来提高性能，从而重新降低加权因子的有效值。

经常出现的词语。有鉴于此，我们引入的交叉熵。

自由地认为它也取决于上下文词语。

结果为

$$J' = \sum_{i,j} f(X_{i(j)})\left(w_i^T\tilde{w}_j - \log X_{i(j)}\right)^2、 \tag{16}$$

[1]，相当于我们之前推导出的公式 (8)中的成本函数。

3.2 模型的复杂性

从公式 (8) 和加权函数 $f(X)$ 的显式可以看出，模型的计算复杂度取决于矩阵 X 中非零元素的数量。由于非零元素数总是少于矩阵的总元素数，因此模型的计算复杂度不会低于 (V^2)。乍一看，这似乎比基于浅层窗口的方法有了很大的改进，因为浅层窗口会随着语料库规模 C 的增大而增大。然而，典型的语料库有成千上万个单词，因此 V^2 可以达到数千亿，这实际上比大多数语料库要大得多。因此，确定是否可以对 $V(2)$ 中的非零元素数量进行更严格的限制非常重要。

X.

为了对 X 中非零元素的数量做出具体说明，有必要对词的共现分布做出一些假设。特别是，我们将假设词 i 与词 j 的共现次数 $X_{(i)(j)}$ 可以用幂律来模拟词对频率等级的函数、$r_{(i)(j)}$：

$$X_{(i)(j)} = \frac{k}{(r_{ij}))^{\alpha}}. \tag{17}$$

有效的补救办法是尽量减少

[1]我们还可以在公式 (16) 中加入偏差项。

语料库中的单词总数与共同出现矩阵 X 中所有元素的总和成正比、

$$|C| \sim \sum_{ij} X_{ij} = \sum_{r=1}^{|X|} k_r = H_{|X|,\alpha} \quad , \qquad (18)$$

其中，我们将最后一个总和改写为

广义谐波数 $H_{n,m}$ 上

总和的每一极限 X 是最大频数秩，与矩阵 X 中的非零元素数相吻合。这个数字也等于公式 (17) 中 r 的最大值，使得 $X_{ij} \geq 1$，即 $X = k(1/\alpha)$。因此，我们可以将公式 (18) 写成 ||

$$|C| \sim |X|^{\alpha} \quad H_{|X|,\alpha}. \qquad (19)$$

我们感兴趣的是，当两个数都很大时，X 与 C 的关系如何；因此，我们可以自由扩展方程右边的大 X。为此目的，我们使用了属调和数展开法（Apostol，1976 年）、

$$H_{x,s} = \frac{x^{1-s}}{1-s} + \zeta(s) + O(x^{-s}) \quad \text{if } s > 0, s \neq 1 \text{、} \qquad (20)$$

给出、

$$|C| \sim \frac{|X|}{1-\alpha} + \zeta(\alpha)|X|^{\alpha} + O(1) \text{、} \qquad (21)$$

其中 $\zeta(s)$ 是黎曼zeta 函数。在 X 较大的极限情况下，公式 (21) 右侧的两个项中只有一项是相关的，至于是哪一项，取决于 $\alpha > 1$、

$$|X| = \begin{cases} O(|C|) & \text{如果 } \alpha < 1\text{、} \\ O(|C|^{1/\alpha}) & \text{如果 } \alpha > 1\text{。} \end{cases} \qquad (22)$$

在本文研究的语料库中，我们发现 $X_{(i)(j)}$ 可以用公式 (17) 很好地建模，其中 α 为 1.25。 = 1.25.在这种情况下，我们可以得出 $|X| = O(|C|^{0.8})$. 因此，我们得出结论，该模型的复杂性比最坏的情况（V^2）要好得多，事实上，它比在线的 O 基于窗口的在线方法要好一些，后者的缩放比例为 $O(|C|)$。

4 实验

4.1 评估方法

我们在 Mikolov 等人（2013a）的单词类比任务、（Luong 等人，2013 年）中描述的各种单词相似性任务以及 CoNLL-2003 共享基准上进行了实验。

表 2：单词类比任务的结果，以准确率百分比表示。带下划线的分数是相似大小模型组内的最佳分数；粗体分数是总体最佳分数。HPCA 向量可公开获取[2]；(i)vLBL 结果来自 (Mnih et al., 2013)；跳格 (SG) 和 CBOW 结果来自 (Mikolov et al., 2013a,2013a) 。(Mikolov et al., 2013a,b)；我们训练了 SG† 和 CBOW†使用 word2vec 工具[3]。详见正文有关 SVD 模型的详细信息和说明。

模型	尺寸	尺寸	Sem.	Syn.	总计
ivLBL	100	1.5B	55.9	50.1	53.2
HPCA	100	1.6B	4.2	16.4	10.8
GloVe	100	1.6B	67.5	54.3	60.3
SG	300	1B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
vLBL	300	1.5B	54.2	64.8	60.0
ivLBL	300	1.5B	65.2	63.0	64.0
GloVe	300	1.6B	80.8	61.5	70.3
SVD	300	6B	6.3	8.1	7.3
SVD-S	300	6B	36.7	46.6	42.1
SVD-L	300	6B	56.6	63.0	60.1
CBOW†	300	6B	63.6	67.4	65.7
SG†	300	6B	73.0	66.0	69.1
GloVe	300	6B	77.4	67.0	71.7
CBOW	1000	6B	57.3	68.9	63.7
SG	1000	6B	66.1	65.1	65.6
SVD-L	300	42B	38.4	58.2	49.2
GloVe	300	42B	**81.9**	**69.3**	**75.0**

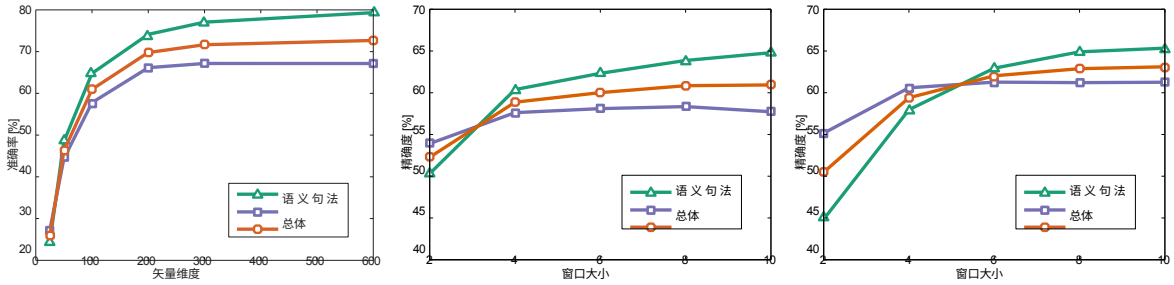
词语类比。词语类比任务包括 "*a* 对 *b* 就像 *c* 对 ?" 这样的问题。该数据集包含 19,544 个此类问题。分为语义子集和句法子集。语义问题通常是类比问题这就是 NER 的数据集（Tjong Kim Sang 和 De Meulder，2003 年）。

句法问题通常是关于动词时态或形容词形式的类比，如 "_雅典之于希腊就像柏林之于？句法问题通常是关于动词时态或形容词形式的类比，例如 "dance is to dancing as fly is to ?"。为了正确回答问题，模型应该唯一地识别出缺失的术语，只有完全对应才算正确匹配。我们在回答 "*a* 与 *b* 的对应关系就像 *c* 与 ?的对应关系一样 "这个问题*时*，根 据 余弦相似度找出其代表词 w_d与 w_b最接近的词 *d*　$w_a + w_{(c)(}$[4)]。

[2]http://lebret.ch/words/

[3]http://code.google.com/p/word2vec/

[4]Levy 等人（2014 年）引入了一种乘法类比评估方法 3COSMUL，并报告说在



(a) 对称背景　　(b) 对称背景　　(c) 非对称环境

图 2：类比任务的准确率与向量大小和窗口大小/类型的函数关系。所有模型都是在 60 亿 token 语料库上训练的。在（a）中，窗口大小为 10。在（b）和（c）中，向量大小为 100。

词语相似性虽然类比任务是我们的主要关注点，因为它可以测试有趣的向量空间子结构，但我们也在表 3 中对各种词语相似性任务对我们的模型进行了评估。这些任务包括 WordSim-353（Finkelstein 等人，2001 年）、MC（Miller 和 Charles，1991 年）、RG（Rubenstein 和 Goodenough，1965 年）、SCWS（Huang 等人，2012 年）和 RW（Luong 等人，2013 年）。

命名实体识别。用于 NER 的 CoNLL-2003 英语基准数据集是路透社新闻通讯文章中的文档合集，其中标注了四种实体类型：人物、地点、组织和杂项。我们在 CoNLL-03 的训练数据上训练模型，并在三个数据集上进行测试：1）ConLL-03 测试数据；2）ACE Phase 2 (2001-02) 和 ACE-2003 数据；3）MUC7。

正式运行测试集。我们采用了 BIO2 注释标准以及（Wang 和 Manning，2013 年）中描述的所有预处理步骤。我们使用了斯坦福 NER 模型标准分布（Finkel 等人，2005 年）中的离散特征综合集。CoNLL- 2003 训练数据集共生成了 437 905 个离散特征。此外，我们还为五字上下文的每个单词添加了 50 维向量，并将其用作连续特征。以这些特征为输入，我们训练了一个条件随机场（CRF），其设置与（Wang 和 Manning，2013 年）的 CRF (join）模型完全相同。

4.2　语料库和训练细节

我们在五个不同规模的语料库上训练了我们的模型：2010 年维基百科语料库，其中有 10 亿个词块；2014 年维基百科语料库，其中有 16 亿个词块；Gigaword 5，其中有 43 亿个词块；Gigaword5+ Wikipedia2014 组合，其中有 5 亿个词块。

类比任务。这个数字是在数据集的一个子集上进行评估的，因此未列入表 2。在几乎所有实验中，3COSMUL 的表现都不如余弦相似度。

有 60 亿个词块；以及来自 Common Crawl[5]的 420 亿个词块的网络数据。我们使用斯坦福 kenizer 对每个语料库进行标记化和小写处理，建立了一个包含 40 万个最常见词汇的词汇表[6]，然后构建了一个共同出现次数矩阵 X。在构建 X 时，我们必须选择语境窗口的大小，以及是否区分左语境和右语境。我们将探讨这些选择的效果。在所有情况下，我们都使用递减加权函数，因此相距 d 个词的词对对总计数的贡献率为 1/d。这是考虑到相距甚远的词对所包含的词与词之间关系的相关信息较少这一事实的一种方法。

在所有实验中，我们将 $x_{\max}=$ 设置为 100、$\alpha=3/4$，并使用 AdaGrad（Duchi 等人，2011 年）训练模型，从 X 中随机抽样非零元素，初始学习率为 0.05.对于小于 300 维的向量，我们进行 50 次迭代，反之则进行 100 次迭代（有关收敛速率的更多详情，请参见第 4.6 节）。除非另有说明，我们使用左侧十个词和右侧十个词的上下文。

该模型生成两组词向量、W 和 $\tilde{W}$。当 X 对称时，W 和 $\tilde{W}$ 是等价的，只有初始化的不同；两组向量的性能应该相当。另一方面，有证据表明，对于某些类型的神经网络，训练网络的多个实例，然后合并结果，有助于减少过拟合和噪音，并普遍改善结果（Ciresan 等人，2012 年）。有鉴于此，我们选择使用

[5]为了证明模型的可扩展性，我们还在一个更大的第六个语料库上对其进行了训练，该语料库包含 840 亿个网络数据，但在这种情况下，我们没有对词汇进行小写，因此结果不具有直接可比性。

[6]对于在 Common Crawl 数据上训练的模型，我们使用了约 200 万字的更大词汇量。

和 $W\tilde{W}$ 作为我们的词向量。这样做通常会略微提高性能，其中语义类比任务的性能提升最大。

我们将其与多种最先进模型的公开结果、我们自己使用 word2vec 工具得出的结果以及几种使用 SVD 的基线结果进行了比较。使用 word2vec，我们在 60 亿 token 语料库（维基百科 2014+ Giga-词 5）上使用了由前 400,000 个最常用词组成的词汇量和 10 个上下文窗口大小的连续词袋（CBOW(†)）模型。我们使用了 10 个负样本，我们将在第 4.6 节中证明这对该语料库来说是个不错的选择。

对于 SVD 基线，我们生成了一个截断矩阵 X_{trunc}，该矩阵只保留了每个词出现频率最高的 10,000 个词的信息。这一步是许多基于矩阵因式分解方法的典型步骤，因为额外的列会带来不成比例的零条目，而且这些方法的计算成本也很高。

该矩阵的奇异向量构成了基准 "SVD"。我们还评估了两个相关的基线：在 "SVD-S "中，我们将 SVD 的 $\sqrt{X_{\text{trunc}}}$ 和 "SVD-L"，其中我们用 SVD 的 $\log(1+X_{\text{trunc}})$。这两种方法都有助于压缩 X 中原本很大的取值范围。

4.3 研究结果

我们在 Ta- ble 2 中展示了单词类比任务的结果。GloVe 模型的表现明显优于其他基线模型，通常在使用较小的向量大小和较小的语料库时也是如此。我们使用 word2vec 工具得出的结果比之前公布的大多数结果都要好一些。这是由多种因素造成的，包括我们选择使用负采样（通常比分层软最大值效果更好）、负采样的数量以及语料库的选择。

我们证明，该模型可以轻松地在 420 亿 token 的大型语料库上进行训练，并相应地大幅提升性能。我们注意到，增加语料库规模并不能保证其他模型的结果得到改善，这一点可以从 SVD- 模型的性能下降中看出。

表 3：单词相似性任务的斯皮尔曼等级相关性。所有向量均为 300 维。CBOW* 向量来自 word2vec 网站。
不同之处在于它们包含短语向量。

模型	大小	WS353	MC	RG	SCWS	RW
SVD	6B	35.3	35.1	42.5	38.3	25.6
SVD-S	6B	56.5	71.5	71.0	53.6	34.7
SVD-L	6B	65.7	72.7	75.1	56.5	37.0
CBOW†	6B	57.2	65.6	68.2	57.0	32.5
SG†	6B	62.8	65.2	69.7	58.1	37.2
GloVe	6B	65.8	72.7	77.8	53.9	38.1
SVD-L	42B	74.0	76.4	74.1	58.3	39.9
GloVe	42B	**75.9**	**83.6**	**82.9**	**59.6**	**47.8**
CBOW*	100B	68.4	79.6	75.4	59.4	45.5

在这一更大的语料库中，SVD 模型的权重也有所提高。这种基本的 SVD 模型不能很好地扩展到大型语料库，这一事实进一步证明了我们的模型中提出的加权方案的必要性。表 3 显示了五个不同词语相似性数据集的结果。首先对词汇表中的每个特征进行归一化处理，然后计算余弦相似度，从而从词向量中获得相似度得分。我们计算斯皮尔曼秩相关系数。
的 SVD 等级相关系数。CBOW* 表示 vec-

我们使用了 word2vec 网站上的词向量和短语向量，这些词向量和短语向量是在 10 亿字的新闻数据中训练出来的。GloVe 的表现优于它，而使用的语料规模还不到它的一半。

表 4 显示了基于 CRF 模型的 NER 任务的结果。如果在 dev 集上迭代 25 次仍无改进，则 L-BFGS 训练终止。除此之外，所有配置均与 Wang 和 Manning（2013 年）使用的配置相同。标为 "*离散*" *(Discrete* ）的模型是基线模型，使用的是斯坦福 NER 模型标准分布的离散特征综合集，但没有词向量特征。除了前面讨论的 HPCA 和 SVD 模型外，我们还与 Huang 等人（2012 年）（HSMN）和 Collobert 与 Weston（2008 年）（CW）的模型进行了比较。我们使用 word2vec 工具[8]训练了 CBOW 模型。在所有评估指标上，GloVe 模型都优于所有其他方法，但在 CoNLL 测试集上，HPCA 方法略胜一筹。我们的结论是，GloVe 向量在下游 NLP 任务中是有用的，正如最初

[7]我们还研究了其他几种转换 X 的加权方案；我们在此报告的方案表现最佳。许多加权方案（如 PPMI）破坏了 X 的稀疏性，因此无法用于大型词汇表。
在词汇量较小的情况下，这些信息理论转换
在词语相似性测量中，这些信息理论转换确实很有效、

[8]我们使用了与上述相同的参数，但在单词类比任务中，这

些参数的表现非常糟糕。

我们发现 5 个负样本的效果略好于 10 个样本。

表 4：使用 50d 向量的 NER 任务的 F1 分数。*离散*是没有词向量的基线。我们使用公开的 HPCA、HSMN 和 CW 向量。详见正文。

模型	偏差	测试	ACE	MUC7
离散	91.0	85.4	77.4	73.4
SVD	90.8	85.7	77.3	73.7
SVD-S	91.0	85.5	77.6	74.3
SVD-L	90.5	84.8	73.6	71.5
HPCA	92.6	**88.7**	81.7	80.7
HSMN	90.5	85.7	78.7	74.7
CW	92.2	87.4	81.7	80.2
CBOW	93.1	88.2	82.2	81.1
GloVe	**93.2**	88.3	**82.9**	**82.2**

图里安等人，2010）中的神经向量所示。

4.4 模型分析：向量长度和上下文大小

在图 2 中，我们展示了改变向量长度和上下文窗口的实验结果。一个上下文窗口延伸到一个词的左边和右边称为对称窗口，一个只延伸到左边的称为不对称窗口。在(a)中，我们观察到维数大于 200 维时，收益会逐渐减少。在(b)和(c)中，我们考察了改变对称和非对称上下文窗口大小的效果。在句法子任务中，小窗口和非对称上下文窗口的性能更好，这与句法信息主要来自直接文本并可能在很大程度上取决于词序的直觉相吻合。另一方面，语义信息通常是非本地信息，窗口越大，捕捉到的语义信息就越多。

4.5 模型分析：语料库规模

图 3 显示了在不同语料库中训练的 300 维向量在单词分析任务中的表现。在句法子任务上，随着语料库规模的增大，性能呈单调增长。这在意料之中，因为更大的语料库通常能产生更好的统计数据。有趣的是，同样的趋势在语义子任务中并不存在，在较小的维基百科语料库中训练的模型比在较大的 Gigaword 语料库中训练的模型表现更好。这可能是由于类比数据集中有大量基于城市和国家的类比，而且维基百科对大多数此类地点都有相当全面的文章。此外，维基百科的

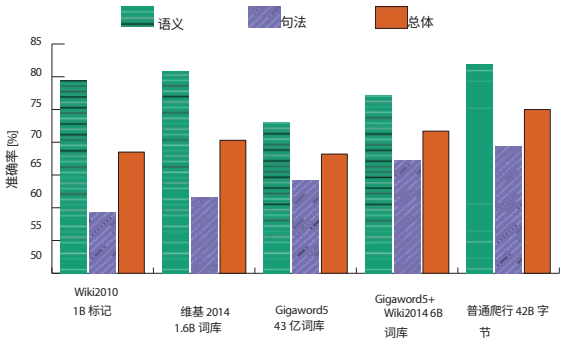


图 3：在不同语料库中训练的 300 维向量的类比任务准确率。

维基百科的词条会不断更新以吸收新知识，而 Gigaword 是一个固定的新闻库，其中的信息可能已经过时，也可能是错误的。

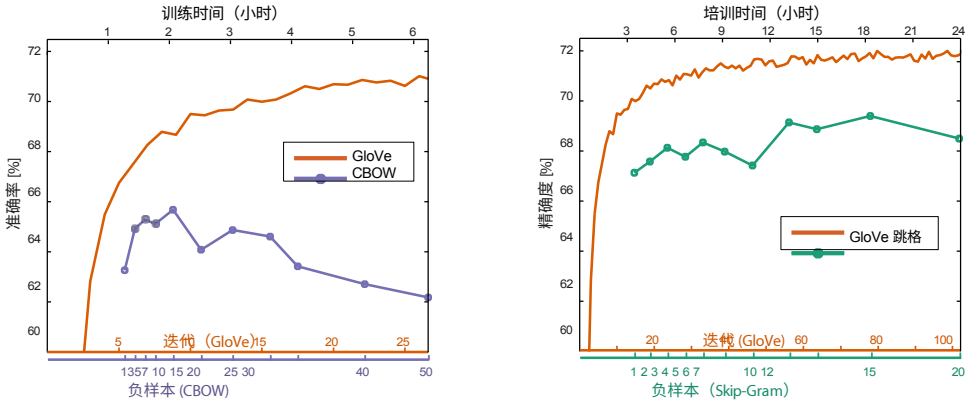
4.6 模型分析：运行时间

总运行时间分为填充 X 和训练模型两部分。前者取决于很多因素，包括窗口大小、词汇量大小和语料库大小。虽然我们没有这样做，但这一步可以很容易地在多台机器上并行化（参见 Lebret 和 Collobert（2014）的一些基准测试）。使用双 2.1GHz 英特尔至强 E5-2658 处理器的单线程，用 10 个词的对称上下文窗口、40 万个词的词汇量和 60 亿个标记语料填充 X 需要大约 85 分钟。在给定 X *的*情况下，训练模型所需的时间取决于向量的大小和迭代次数。对于上述设置下的 300 维向量（并使用上述模型的全部 32 个内核），单次迭代需要 14 分钟。学习曲线图见图 4。

4.7 模型分析：与 word2vec

对 GloVe 和 `word2vec` 进行严格的定量比较非常复杂，因为存在许多对性能有很大影响的参数。我们通过将向量长度、上下文窗口大小、cor- pus 和词汇量大小设置为上一小节中提到的配置，来控制我们在第 4.4 和 4.5 节中确定的主要变量来源。

剩下需要控制的最重要变量是训练时间。对于 GloVe，重要的参数是训练迭代次数。对于 `word2vec`，显而易见的选择是训练迭代次数。遗憾的是，该代码目前只设计了单次训练：



(a) GloVe vs CBOW　　(b) GloVe vs Skip-Gram

图 4：单词类比任务的总体准确率与训练时间的函数关系，GloVe 的训练时间取决于迭代次数，CBOW 的训练时间取决于负样本的数量（a），而 skip-gram 的训练时间取决于负样本的数量（b）。

(b).在所有情况下，我们都在相同的 6B token 语料库（维基百科 2014+ Gigaword 5）上训练 300 维向量，词汇量同样为 40 万，并使用大小为 10 的对称上下文窗口。

它指定了一个针对单次数据学习的学习计划，因此对多次数据进行修改并非易事。另一种选择是改变负样本的数量。增加负样本可以有效增加模型看到的训练词数量，因此在某些方面类似于额外的历时。

我们将任何未指定的参数设置为其故障值，假定它们接近最佳值，不过我们承认，在进行更全面的分析时应放宽这种简化。在图 4 中，我们绘制了类比任务的总体性能与训练时间的函数关系图。底部的两个 *x 轴*分别表示 GloVe 和 `word2vec` 相应的训练迭代次数和负样本数。我们注意到，如果负样本数量增加到 10 个以上，`word2vec` 的性能实际上会下降。这可能是因为负采样方法不能很好地近似目标概率分布。目标概率分布[9]。

在语料、词汇、窗口大小和训练时间相同的情况下，GloVe 的表现始终优于 `word2vec`。它能以更快的速度获得更好的结果，而且无论速度如何，都能获得最佳结果。

5 结论

最近，分布式单词表征是否最适合从基于计数的单词表征中学习这一问题受到了广泛关注。或基于预测的方法。目前，基于预测的模型获得了大量支持；例如，Baroni 等人（2014 年）认为这些模型在一系列任务中表现更好。在本研究中，我们认为这两类方法在本质上并无显著区别，因为它们都能探究语料库中不为人知的共现统计数据，但基于计数的方法捕捉全局统计数据的效率更具优势。我们构建了一个模型，利用计数数据的这一主要优势，同时捕捉最近基于对数-双线性预测方法（如 `word2vec`）中流行的有意义的线性子结构。GloVe 是一种新的全局对数-双线性回归模型，可用于词表征的无监督学习，在词类比、词相似性和命名实体识别任务中的表现优于其他模型。

致谢

感谢匿名审稿人的宝贵意见。斯坦福大学感谢国防威胁降低局 (DTRA) 根据空军再搜索实验室 (AFRL) 合同编号 FA8650- 10-C-7020 和国防高级研究计划局 (DARPA) 深度探索和文本过滤计划 (DEFT) 提供的支持。FA8650- 10-C-7020 和美国国防部高级研究计划局 (DARPA) 文本深度探索和过滤 (DEFT) 计划的支持，合同编号为 AFRL.FA8750-13-2-0040。FA8750-13-2-0040。本资料中的任何观点、发现、结论或建议均属作者个人观点，并不一定反映 DTRA、AFRL、DEFT 或美国政府的观点。

[9]与此相反，噪声对比估计是一种近似方法，它能通过更多的负样本得到改进。在（Mnih 等人，2013 年）的 Ta- ble 1 中，类比任务的准确率是负样本数量的非递减函数。

和

参考文献

Tom M. Apostol.1976.*解析数论导论*》。解析数论导论》。

Marco Baroni, Georgiana Dinu, and Germa´n Kruszewski.2014.不计算，预测！语境计算与语境预测语义向量的系统比较。In *ACL*.

Yoshua Bengio.2009.为人工智能学习深度架构》。*机器学习的基础与趋势*》。

Yoshua Bengio、Re´jean Ducharme、Pascal Vin- cent 和 Christian Janvin。2003.神经概率语言模型。*JMLR*，3：1137-1155。

John A. Bullinaria and Joseph P. Levy.2007.从词的共同出现统计中提取语义表征：A computational study. *行为研究方法*》，39（3）：510-526。

Dan C. Ciresan, Alessandro Giusti, Luca M. Gam- bardella, and ¨ rgen Schmidhuber.2012.深度神经网络分割电子显微镜图像中的神经元膜。在 *NIPS 上*，第 2852-2860 页。

Ronan Collobert 和 Jason Weston。2008.自然语言处理的统一架构：多任务学习的深度神经网络。In *Proceedings of ICML*, pages 160-167.

Ronan Collobert、Jason Weston、Le´on Bottou、Michael Karlen、Koray Kavukcuoglu 和 Pavel Kuksa。2011.从零开始的自然语言处理（Al- most）》。*JMLR*，12：2493-2537.

Scott Deerwester、Susan T. Dumais、George W. Furnas、Thomas K. Landauer 和 Richard Harshman。1990.通过潜在语义分析编制索引。*美国信息科学学会期刊*》，41.

约翰-杜奇、埃拉德-哈赞、约拉姆-辛格。2011.在线学习和随机优化的自适应子梯度方法。*JMLR*，12。

Lev Finkelstein, Evgenly Gabrilovich, Yossi Ma- tias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin.2001.将搜索置于文本中：概念重温。*第 10 届万维网国际会议论文集*》，第 406-414 页。ACM.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng.2012.通过全局上下文

通过全局上下文和多个单词原型改进单词表征。见 *ACL*。

Re´mi Lebret 和 Ronan Collobert。2014.通过海灵格 PCA 进行词嵌入。见 *EACL*。

Omer Levy、Yoav Goldberg 和 Israel Ramat- Gan。2014.稀疏和显性单词表征中的语言规律性。*CoNLL-2014*。

Kevin Lund 和 Curt Burgess。1996.从词汇共现生成高维语义空间》。*行为研究方法、仪器和计算机*》，28：203-208。

Minh-Thang Luong、Richard Socher 和 Christo-pher D Manning。2013.用递归神经网络更好地代表词的形态。*CoNLL-2013*.

Tomas Mikolov、Kai Chen、Greg Corrado 和 Jef- frey Dean。2013a.向量空间中单词表示的高效估计。In *ICLR Work- shop Papers*.

Tomas Mikolov、Ilya Sutskever、Kai Chen、Greg Corrado 和 Jeffrey Dean。2013b.单词和短语的分布式表示及其构成性。In *NIPS*, pages 3111-3119.

Tomas Mikolov、Wen tau Yih 和 Geoffrey Zweig。2013c.连续空间词表征中的语言规律性。In *HLT-NAACL*.

George A. Miller and Walter G. Charles.1991.语义相似性的语境相关性。*语言与认知过程*》，6（1）：1-28。

Andriy Mnih and Koray Kavukcuoglu.2013.利用噪声对比估计高效学习词嵌入。In *NIPS*.

Douglas L. T. Rohde, Laura M. Gonnerman, and David C. Plaut.2006.基于词法共存的语义相似性改进模型。*ACM 通信*》，8:627-633。

Herbert Rubenstein and John B. Goodenough.1965.同义词的语境相关性。*ACM 通信*，8（10）：627-633.

Fabrizio Sebastiani.2002.自动化文本分类中的机器学习。*ACM Computing Surveys*, 34:1-47.

Richard Socher、John Bauer、Christopher D. Man-ning 和 Andrew Y. Ng。2013.使用合成向量语法进行解析。In *ACL*.

Stefanie Tellex、 Boris Katz、 Jimmy Lin、 Aaron Fernandes 和 Gregory Marton。2003.用于问题解答的段落检索算法的量化评估》（Quanti-tative evaluation of passage retrieval algorithms for question answering ）。 In *Proceedings of the SIGIR Conference on Research and Develop- ment in Informaion Retrieval*.

Erik F. Tjong Kim Sang and Fien De Meul- der.CoNLL-2003 共享任务简介：语言无关的命名实体识别。In *CoNLL-2003*.

Joseph Turian、Lev Ratinov 和 Yoshua Bengio。2010. 单词表征：半监督学习的简单通用方法。*ACL 论文集*，第 384-394 页。

Mengqiu Wang and Christopher D. Manning.2013.非线性深度架构在序列标注中的效果。*第六届国际自然语言处理联合会议（IJCNLP）论文集*》。