基于词嵌入的性别偏见度量对频率的不良依赖

Francisco ValentiniICC（UBA-CONICET）数据挖掘硕士课程（UBA）布宜诺斯艾利斯，阿根廷 ft.valentini@gmail.com

Germán RosatiIDAES学院（UNSAM）布宜诺斯艾利斯，阿根廷 grosati@unsam.edu.ar

Diego Fernandez Slezak
ICC（UBA-CONICET）布宜诺斯艾利斯，阿根廷
dfslezak@dc.uba.ar

Edgar Altszyler
ICC（UBA-CONICET）数据挖掘硕士课程（UBA）阿根廷布宜诺斯艾利斯 ealtszyler@dc.uba.ar

arXiv:2301.00792v1 [cs.CL] 2 Jan 2023

摘要

许多研究使用基于词嵌入的度量来量化文本中的社会偏见和刻板印象。近期研究发现，词嵌入能够捕捉语义相似度，但可能受到词频的影响。在本研究中，我们探讨使用基于词嵌入的偏见量化方法测量女性与男性之间的性别偏见时，频率所产生的影响。我们发现，采用负采样的 Skip-gram 和 GloVe 往往会在高频词中检测到男性偏见，而 GloVe 往往会在低频词中得出女性偏见。我们表明，即使将词随机打乱，这些现象仍然存在。这证明，在未打乱语料库中观察到的基于频率的效应，源于度量本身的属性，而非词语联想。该效应是虚假的且有问题，因为偏差指标应完全取决于词共现，而非单个词的频率。最后，我们将这些结果与基于点互信息的另一种度量所得的结果进行比较。我们发现，该度量并未表现出明显的频率依赖性，尽管在所有频率范围内都略微偏向男性偏见。

1 引言

词嵌入是衡量语料库中词语之间语义接近程度的最常用技术之一。近年来，词嵌入被广泛应用于计算社会科学领域，用于衡量社会偏见和刻板印象（Caliskanetal.，2017年；Garg 等人，2018；Kozlowski 等人，2019年；Lewis 和 Lupyan，2020；Charlesworth 等人，2021）。

出于实际考虑，我们将偏见定义为用于描述群体或事物的语言存在差异的程度。偏见通常通过计算两个上下文组A 和 B 中的词相对于某个目标词 x的平均相似度之差来衡量：

$$\text{Bias}_{\text{WE}} = \underset{a\in A}{\text{mean}} \cos(v_x, v_a) - \underset{b\in B}{\text{mean}} \cos(v_x, v_b), \tag{1}$$

其中， v_i 是词 i 的词嵌入，cos(v_i,v_j)是两个向量之间的余弦相似度。

性别偏见长期以来一直是采用这种方法研究最多的偏见之一。在此背景下， A 和B 通常被定义为性别化名词和代词（Caliskan 等人，2017年；Lewis 和 Lupyan，2020）。一个典型例子是Garg 等人（2018），他们发现，历史语料库中职业的女性与男性偏见，与不同时期各职业女性从业者所占比例相关。

点互信息（PMI）不如词嵌入应用广泛，它是一种衡量词语相似度的指标，也可用于研究偏见（Gálvez 等人，2019年；Aka 等人，2021；Valentini 等人，2021）。Valentini 等人（2021）将基于点互信息的偏见指标定义为

$$\text{Bias}_{\text{PMI}} = \text{PMI}(x, A) - \text{PMI}(x, B),$$

其中

$$\text{PMI}(x, Y) = \log \frac{P(x, Y)}{P(x)P(Y)}.$$

P(x， Y) 表示在预定义词数的窗口内，词 x 与 Y中的任意一个词共现的概率，而 P(x) 和

P（Y）分别表示词x 与 Y中任意一个词出现的概率。Valentini 等人（2021）表明，偏见$_{\text{PMI}}$ 可以表示为

$$\text{Bias}_{\text{PMI}} = \log \frac{P(x|A)}{P(x|B)}. \quad (2)$$

也就是说，偏见$_{\text{PMI}}$ 可以理解为：在 x 中的词出现在 A 中的词的语境下的可能性，比出现在 B中的词的语境下高出多少（以对数尺度表示）。因此，它只捕捉一阶关联，并可通过对文本中的共现计数使用最大似然法计算得出（Valentini 等人，2021）。

近期有一些研究探讨了词频与词嵌入之间的关系。具体而言，即使经过归一化，嵌入似乎仍会编码词频（Schnabel 等人，2015）；向量范数取决于词频（Wilson 和 Schakel，2015）；主成分的前几个方向以不同方式编码频率（Mu 和 Viswanath，2018）；高频词和低频词的向量位于嵌入空间的不同区域（Gong 等人，2018）。

然而，这些研究的结论尚不明确，因为它们并未清楚判定观察到的关联在多大程度上是由语料库的实际属性造成的，又在多大程度上源于嵌入训练中不理想的特性。因此，嵌入与频率之间关系的成因仍有待探明。此外，这一效应在计算社会科学相关应用中的影响（例如偏见量化）尚未得到探索。

我们作出了三项主要贡献。第一，我们表明，使用基于词嵌入的指标进行测量时，频率与性别偏见相关：基于 Skip-gram with negative sampling (SGNS) 和 GloVe 的偏差指标都倾向于在高频词中检测出男性偏见，而 GloVe 平均而言也会在低频词中检测出女性偏见。第二，我们表明，即使语料库中的词元经过随机打乱，基于词嵌入的性别偏见对频率的依赖仍然存在。这证明，对频率的依赖是指标本身造成的假象，也就是说，基于词嵌入的偏差指标可能会错误地编码频率。第三，我们发现，基于 PMI 的性别偏见指标不会呈现这种基于频率的效应，但在所有频率范围内都略微偏向男性偏见。[1]

[1]本文的代码可在 https://github.com/ftvalentini/EmbeddingsBiasFrequency 获取

我们的分析仅限于英语，并基于二元性别观（参见局限性）。

2 频率对性别偏见的影响

我们的目标是研究广泛使用的基于嵌入的指标和另一种基于点互信息的指标中，性别偏见与频率之间的关联。因此，在第一个实验中，我们分析了两个预训练词嵌入中的这种关联：GloVe（Pennington 等人，2014）和采用 SGNS 的 Word2Vec（Mikolov 等人，2013）。

我们通过研究词汇表中不同频率区间内词的偏见分布来实现这一点。我们使用1 式计算偏见，所用的女性和男性上下文词列表来自 Caliskan 等人（2017年），并将其称为女性偏见或性别偏见。对于每个频率区间，我们还计算均值与样本标准差（SD）之比。这些是衡量零假设下各组均值的效应量，即平均而言不存在偏见（Cohen，1988年）。在此，我们将其用作分布偏离零值的标准化幅度。下文将使用这一方法评估性别偏见与频率之间的关联。更多详情请参阅附录B 。

性别偏见与$_{\text{WE}}$ 预训练嵌入以及目标词的词频之间存在明显关联（图1）。GloVe 嵌入呈现出频率与性别偏见之间的单调关系：排名前 $10^{3.5}$的词往往具有男性偏见，且效应量较大；而频率较低的词则平均呈现女性偏见，效应量中到大。SGNS 嵌入中，频率较低的词效应较小且为正向，但在排名前 100 的词中，偏见则大幅转向男性偏见。

尽管已有文献研究了频率与词向量之间的关系（见第1节），这一结果仍令人惊讶：事前看来，我们原本不会预期词语的性别偏见与频率如此强烈地相关，因为词语相似度理应与语义以及训练语料库中的共现关系联系更紧密，而非与词语各自的频率相关。

为验证这一现象，我们使用英文维基百科从头训练 GloVe 和 SGNS 嵌入，并研究性别偏见与词频之间的关联。我们将其与使用 Bias$_{\text{PMI}}$ （公式 2）获得的结果进行比较。

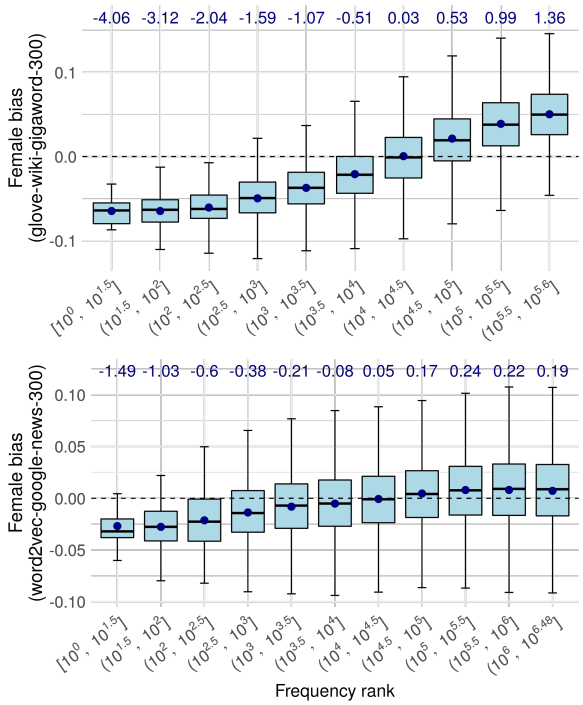


图 1：预训练 GloVe（上图）和采用 SGNS 的 Word2Vec（下图）中，女性偏见分布与词频率排名的关系。根据词的对数尺度排名将其分组，每组对应一个区间，因此频率最高的词位于最左侧区间，频率较低的词位于最右侧区间。由于预训练嵌入未提供原始频率，我们使用频率排名。蓝点表示均值，蓝色数值表示效应量（均值与标准差之比）。由于每组嵌入所使用的语料库、词汇表和训练方法不同，这些图在任一坐标轴上都不可直接比较。

2.1 比较 Bias_{WE} 与 Bias_{PMI}

方法与数据 我们使用 Bias_{PMI} 和 Bias_{WE} 测量 2021年英文维基百科词汇表中词语的性别偏见，并评估其与词频之间的关联。我们训练 SGNS 和 GloVe 向量以计算 Bias_{WE}，而使用语料库中的频次统计来计算 Bias_{PMI}。有关语料库和方法的详细信息，请分别参阅附录A 和 B 。

结果预训练嵌入中偏见$_{\text{WE}}$ 与频率之间的关系（图 1）在从头开始训练嵌入时，在定性上仍然成立（图 2的上图和中图）。GloVe 嵌入中，女性偏见与频率呈负相关；而在 SGNS 中，我们发现高频词表现出平均程度的男性偏见，且效应量中等至较大。

使用 Bias_{PMI} 时，不存在任何频率区间在平均意义上存在强烈偏见（图2的下图）。不过，整体略微偏向男性偏见，所有频率范围的效应量都略呈负值。此外，随着目标词频率降低，偏见的变化幅度往往会增加：这是因为低频词的点互信息通常较高且噪声较大（Jurafsky 和 Martin，2009年）。

仅凭这项分析，还不足以断定频率对基于词嵌入的偏差指标的影响是由嵌入造成的虚假现象。高频词仍有可能确实比低频词带有更强的男性偏见，这是由二阶或更高阶关联导致的，因而会产生图2上方和中间面板所示的那类结果。为评估这一点，我们开展了以下研究。

2.2 对频率的不理想依赖

方法与数据 我们创建了五个随机打乱、相互独立的维基百科语料库版本，其中词元在文本中的位置均经过随机排列。在这些语料库中，词语的频率保持不变，但由于共现完全随机，它们失去了上下文。我们在每个语料库中使用 Bias_{WE} 和 Bias_{PMI} 估算偏见，并将五个值的平均值视为每个词语的性别偏见。我们分析了这种情况下性别偏差指标与频率之间的关系。通过打乱语料库中的词语，上下文变得毫无意义，因此，在这种情况下发现的任何偏见与频率之间的关联，都只能由词语的频率解释。我们强调，在不存在偏见的情况下，指标却检测出偏见，这是有问题且不可取的。

结果在这一受控实验设置中，偏见$_{\text{WE}}$ 与目标词频呈现出较强的关联（图3）：对于 SGNS 和 GloVe，平均男性偏见都会随着频率增加而增强，并且从大约频率 10^4 开始，效应量较大。使用 GloVe 测量时，低频词平均表现出女性偏见；而使用 SGNS 时，它们往往表现出轻微的男性偏见，且效应量较小。

相反，在打乱后的语料库中使用点互信息衡量偏见，并不会呈现出明显的频率依赖性。所有频率下的平均偏见大致保持不变，效应量略为负值；

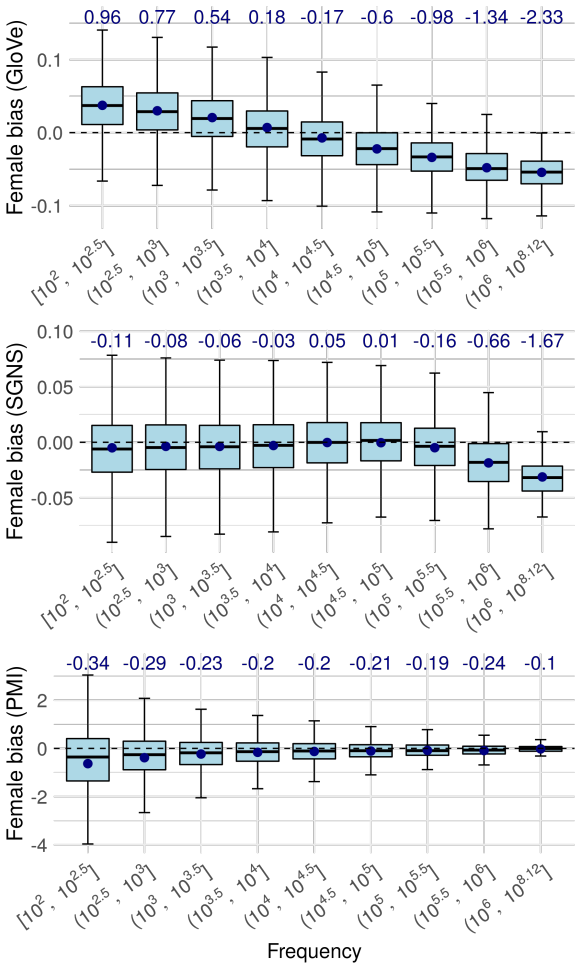


图2：维基百科中的女性偏见与频率。使用 GloVe（上图）、SGNS（中图）和 $_{\text{PMI}}$（下图）计算偏见，分别得到 Bias_{WE}、Bias_{WE} 和 Bias_{PMI}。词汇表中的词按其频率以对数尺度分组。蓝点表示均值，蓝色数值表示效应量（均值与标准差之比）。

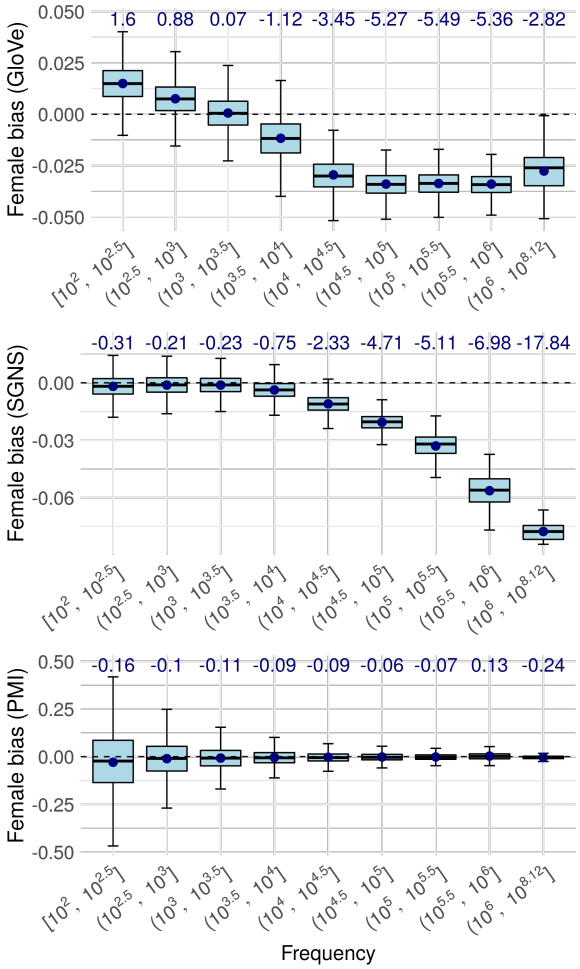


图3：打乱后的维基百科中的女性偏见与频率。每个词的偏见是五次估计的平均值，每次估计对应一次打乱，共进行了五次打乱。词按其频率分组。蓝点表示均值，蓝色数值表示效应量（均值与标准差之比）。

也就是说，在所有频率下都略微偏向男性偏见。

3 讨论与结论

在这项工作中，我们揭示了基于词嵌入之间余弦相似度的性别偏见指标中存在一种虚假的、基于频率的失真。基于 SGNS 和 GloVe 的性别偏见指标往往会在高频词中检测到男性偏见，而 GloVe 平均而言也会在低频词中呈现女性偏见。

为了确定这种效应是否确实是基于嵌入的指标所导致的不良产物，我们评估了打乱的语料库中性别偏见与频率之间的关系；在这些语料库中，词失去了上下文，但保留了各自的频率。结果显示，对频率的依赖是由指标造成的，并非源自文本的实际属性。这表明，即使不存在偏见，常用的性别偏见测量方法也可能检测出偏见。此外，我们发现，另一种基于点互信息的偏见指标没有表现出明显的频率依赖性，尽管它略微倾向于男性偏见。

根据这些结果，我们认为基于点互信息的偏见指标优于基于嵌入的指标，这进一步体现了其在可解释性和假设检验方面的优势（Valentini 等人，2021）。然而，由于点互信息仅能捕捉一阶关联，无法捕捉同义词，因此可能需要纳入若干与上下文词相关的词语，以

便测量某些偏见。

在大型语料库中，男性名词和代词通常比女性名词和代词更常见（Twenge 等人，2012；Gálvez 等人，2019年）。例如，在维基百科语料库中，他出现了 1180 万次，而她的频率为 350 万次（其他性别化上下文词的频率请参见附录B ）。

男性和女性语境的频率差异本身就是一种偏见，可以通过统计词语出现次数来衡量。相比之下，我们研究的偏见指的是男性和女性实体被描绘时所处的刻板印象语境，它应当独立于单个词的频率。

打乱词语顺序后，与女性和男性上下文词的语境相关的偏见会消失，但频率差异仍然存在。我们认为，偏差指标捕捉的是女性和男性上下文词的频率差异。对于基于嵌入的指标，现有证据表明，嵌入除了编码语义外，还编码了词频，这支持了这一假设。

我们认为，随机打乱实验具有足够的普适性，能够表明只要上下文词的频率不同，即使换用其他词表、偏见类型和领域，频率效应依然存在。这一结果十分重要，因为使用嵌入测量偏见时，往往会忽略上下文词的频率。

我们的发现对偏见测量应用具有重要意义，因为当涉及的词频率差异很大时，这些发现使人们对广泛使用的偏差指标的可靠性产生了怀疑。我们认为，应投入更多精力，设计不受这一缺陷影响的新偏见检测方法。

局限性

我们使用了性别偏见文献中常见的上下文词集合。这些词体现了二元性别观，使其他性别表征无法纳入偏见测量。此外，我们只关注英文维基百科语料库，并未将这些方法应用于其他领域的语料库，而那些语料库中的性别偏见分布可能有所不同。

我们报告了使用默认超参数得到的结果。这旨在模拟计算社会科学文献中常见的实验设置。超参数保留默认值，因为偏见不存在真实标注，也就是说，没有标注指出词语的偏见程度。

本研究采用的方法可以推广到其他语言、其他类型的偏见以及其他语料库。我们希望后续研究也能评估这些情境下基于频率的失真，以及超参数选择的影响。

参考文献

Osman Aka、Ken Burke、Alex Bauerle、Christina Greer 和 Margaret Mitchell。2021 年。在缺少真实标注的情况下衡量模型偏差。载于*2021 年 AAAI/ACM 人工智能、伦理与社会会议论文集*。ACM。

Aylin Caliskan、Joanna J. Bryson 和 Arvind Narayanan。2017年。从语言语料库自动推导出的语义包含类人偏见。*Science*，356(6334):183–186。Tessa ES Charlesworth、Victor Yang、Thomas C Mann、Benedek Kurdi 和 Mahzarin R Banaji。2021。自然语言中的性别刻板印象：词嵌入显示，在超过6500万词的儿童和成人语言语料库中，性别刻板印象具有稳健的一致性。*Psychological Science*，32(2):218–240。Jacob Cohen。1988年。行为科学统计功效分析。Routledge。Ramiro H. Gálvez、Valeria Tiffenberg 和 Edgar Altszyler。2019年。电影中性别与智力能力之间刻板印象关联的半个世纪。*Sex Roles*，81(9):643–654。Nikhil Garg、Londa Schiebinger、Dan Jurafsky 和 James Zou。2018。词嵌入量化了100年来的性别和族群刻板印象。美国国家科学院院刊，115(16):E3635–E3644。Chengyue Gong、Di He、Xu Tan、Tao Qin、Liwei Wang 和 Tie-Yan Liu。2018。Frage：与频率无关的词表示。载于神经信息处理系统进展，第31卷。Curran Associates, Inc.。Daniel Jurafsky 和 James H. Martin。2009年。语音与语言处理（第*2*版）。Prentice-Hall, Inc.，美国。Austin C. Kozlowski、Matt Taddy 和 James A. Evans。2019年。文化几何：通过词嵌入分析阶级的含义。*AmericanSociological Review*，84(5):905–949。Omer Levy、Yoav Goldberg 和 Ido Dagan。2015。借鉴词嵌入的经验改进分布相似性。计算语言学协会学报，3:211–225。

Molly Lewis 和 Gary Lupyan。2020。性别刻板印象反映在25种语言的分布结构中。*Nature Human Behaviour*，4(10):1021–1028。

Tomas Mikolov、Ilya Sutskever、Kai Chen、Greg S Cor- rado 和 Jeff Dean。2013 年。词语和短语的分布式表示及其组合性。载于神经信息处理系统进展，第 26 卷。Curran Associates, Inc.

Jiaqi Mu 和 Pramod Viswanath。2018 年。去除顶层：一种简单有效的词表示后处理方法。载于国际学习表征会议。

Jeffrey Pennington、RichardSocher 和 Christopher D. Manning。2014 年。GloVe：用于词表示的全局向量。载于自然语言处理实证方法会议（*EMNLP*），第 1532–1543 页。

Radim ˇReh ̊uˇrek 和 Petr Sojka。2010 年。面向大型语料库的主题建模软件框架。载于*LREC 2010*自然语言处理框架新挑战研讨会论文集，第 45–50 页，马耳他瓦莱塔。ELRA。
http://is.muni.cz/publication/884893/en。

Tobias Schnabel、Igor Labutov、David Mimno 和 Thorsten Joachims。2015 年。无监督词嵌入的评估方法。载于*2015* 年自然语言处理实证方法会议论文集，第 298–307 页，葡萄牙里斯本。计算语言学协会。

Jean MTwenge、W Keith Campbell 和 Brittany Gentile。2012 年。美国书籍中的男性和女性代词使用反映了女性地位，1900–2008。*Sex Roles*，67(9):488–493。

Francisco Valentini、Germán Rosati、Damián Blasi、Diego Fernandez Slezak 和 Edgar Altszyler。2021 年。关于文本中偏见指标的解读与意义：一种基于点互信息的方法。*arXiv* 预印本 *arXiv:2104.06474*。

Benjamin J Wilson 和 Adriaan MJ Schakel。2015 年。词嵌入的受控实验。*arXiv* 预印本 *arXiv:1510.02675*。

A 语料库

我们使用2021年4月的维基百科转储[2]，并删除少于50个词元的文章。我们移除非字母数字符号，并进行句子切分。预处理后，语料库约包含17亿个词元和7810万篇文档（句子）。

[2]https://archive.org/download/enwiki-20210401

B 方法

我们使用性别化名词和代词来衡量女性与男性性别（Caliskan 等人，2017年；Lewis 和 Lupyan，2020），具体而言，A=*{*女性，女人，女孩，姐妹，她，她，她的，女儿*}*和 B=*{*男性，男人，男孩，兄弟，他，他，他的，儿子*}*。

表1和2 显示这些词各自在预处理后的维基百科语料库中的频率。

Word	频率
her	3,720,408
she	3,517,570
女儿	294,043
女性	282,159
女人	236,954
姐妹	179,511
girl	141,616
hers	5,706

表1：维基百科语料库中女性语境词的频率

Word	频率
he	11,815,189
his	9,603,118
him	1,811,552
son	541,828
man	443,881
兄弟	287,544
male	181,471
boy	124,326

表2：维基百科语料库中男性语境词的频率

我们剔除了出现次数少于 100 次的词，得到包含 222,144 个词的词汇表。表格3 展示了这些词按出现频率划分的分布，其中不包括女性和男性上下文词。

频率	词数
$[10^2, 10^{2.5}]$	116,340
$(10^{2.5}, 10^3]$	54,187
$(10^3, 10^{3.5}]$	26,617
$(10^{3.5}, 10^4]$	13,144
$(10^4, 10^{4.5}]$	6,579
$(10^{4.5}, 10^5]$	3,255
$(10^5, 10^{5.5}]$	1,448
$(10^{5.5}, 10^6]$	441
$(10^6, 10^{8.12}]$	117

表3：维基百科语料库中各频率范围内的词数

在第 2 节中，我们使用了在 Wikipedia 2014 和 Gigaword 5 上训练的预训练 GloVe 词嵌入（Pennington 等人，2014），以及在 Google 新闻上训练的 Word2Vec SGNS 词嵌入（Mikolov 等人，2013），二者的维度均为 300。

第2.1 节和2.2节中使用的所有方法（GloVe、SGNS 和点互信息）均采用 10 的窗口大小，并在将语料库处理为词-上下文对之前移除词表外词元（Levy 等人，2015）。

对于 SGNS，我们使用 Gensim 库中的 Word2Vec 实现，ˇReh ̊uˇrek 和 Sojka，2010），并采用默认超参数。GloVe 使用Pennington 等人 （2014）的实现进行训练，迭代 100 次。

对于点互信息，我们使用 GloVe 模块统计共现（Pennington 等人，2014），并将平滑参数 ϵ 设为 0.01，这样即使目标词与任何上下文词之间都没有共现，也能进行计算。

所有计算均在一台台式计算机上完成，配备 4 核英特尔 Core i5-4460 处理器 @ 3.20GHz 和 32 GB 随机存取存储器。使用 GloVe 训练时，每次迭代约需 30 分钟；使用 SGNS 训练时，每个轮次约需 2 小时。