**A SIMPLE BUT TOUGH-TO-BEAT BASELINE FOR SENTENCE EMBEDDINGS**

ABSTRACT

计算单词嵌入的神经网络方法的成功激发了生成较长文本片段（如句子和段落）语义嵌入的方法。令人惊讶的是，Wieting等人（ICLR'16）的研究表明，这种复杂的方法在领域外（转移学习）环境下，通过对单词嵌入进行温和的再训练和基本线性回归的简单方法，其性能优于其他方法（Ganitkevitch等人，2013年）等的方法。需要使用大量标记的数据集（如释义数据库）进行再培训。

本文进一步指出，以下完全无监督的句子嵌入是一个令人生畏的基线：使用维基百科等未标记语料库中常用的方法之一计算的单词嵌入，用单词向量的加权平均来表示句子，然后使用PCA/SVD对其进行一点修改。在文本相似性任务中，这种加权方法的性能提高了约10%到30%，优于包括RNN和LSTM在内的复杂的监督方法，甚至改进了Wieting等人的嵌入。这个简单的方法应该作为将来的基线，特别是当标记的训练数据稀少或不存在时。

本文还从理论上解释了上述方法的成功之处，即在Arora等人的模型基础上，对潜在变量生成模型进行了简单的扩展。（TACL'16）使用新的“平滑”术语，允许出现上下文以外的单词，以及类似和并非所有上下文中的单词的高概率。

1 INTRODUCTION

使用不同方法计算的单词嵌入是自然语言处理（NLP）和信息检索（IR）的基本构件。它们捕捉了单词之间的相似性（例如，Bengio等人，2003年；Collobert&Weston，2008年；Mikolov等人，2013a；Pennington等人，2014年）。最近的工作试图计算捕捉单词序列（短语、句子和段落）语义的嵌入，方法从简单的单词向量的附加组合到复杂的结构，如卷积神经网络和递归神经网络（例如，Iyyer等人，2015；Le&Mikolov，2014；Kiros等人，2015年；Socher等人，2011年；Blunsom等人，2014年；Tai等人，2015年；Wang等人，2016年）。最近，（Wieting et al.，2016）从标准单词嵌入开始，在释义对数据集（PPDB）的监督下对其进行修改，通过训练简单的单词平均模型来构建句子嵌入，从而学习了通用的释义句子嵌入。这种简单的方法比各种各样的方法在文本相似性任务上有更好的性能，并且可以作为文本分类任务的良好初始化。然而，来自释义数据集的监督似乎是至关重要的，因为它们报告初始单词嵌入的简单平均值并不能很好地工作。

在这里，我们给出了一种新的句子嵌入方法，这种方法非常简单：只需计算句子中词向量的加权平均值，然后去除平均向量在其第一个奇异向量上的投影（“公共成分去除”）。这里，单词w的权重是

其中a是一个参数，p（w）是（估计的）单词频率；我们称之为平滑反频率（SIF）。该方法在各种文本相似性任务上的性能明显优于未加权平均值，并且在大多数任务上甚至优于（Wieting等人，2016）中测试的一些复杂的监督方法，包括一些RNN和LSTM模型。该方法非常适合于领域自适应设置，即在各种语料库上训练的词向量用于计算不同测试平台上的句子嵌入。该算法对加权方案也具有较强的鲁棒性：使用不同语料库中的词频率估计不会影响性能；参数a的范围较宽，可以达到接近最优的效果，而参数a的范围较宽，可以比未加权平均值有显著的提高。

当然，如果将“句子”视为“文档”，并合理地假设句子通常不包含重复的单词，则这种SIF重新称重会让人联想到信息检索中的TF-IDF重新称重（Spark Jones，1972；Robertson，2004）。这种重新加权（或相关的想法，如从词汇表中删除频繁的单词）是一个很好的经验法则，但在嵌入单词的设置中没有理论依据。

本论文提供了使用句子生成模型重新称重的理论依据，这是对RandomWalk on语篇生成模型的简单修改（Arora等人，2016）。文中指出，该模型在理论上隐含着一个句子嵌入，即所有单词嵌入的简单平均值。

我们对这个理论模型进行了修正，其动机是经验观察到，大多数的单词嵌入方法，因为它们试图利用向量内积来捕获单词的共现概率，最终给频繁的单词赋予大向量，给单词对赋予不必要的大内积，仅仅是为了符合经验主义的观察，即文字有时出现在文档的上下文之外。这些异常导致词向量的平均值沿着语义上无意义的方向有巨大的分量。我们对生成模型（Arora等人，2016）的修改允许“平滑”项，然后最大似然计算导致我们的SIF重新加权。

有趣的是，这个理论推导的SIF在我们的环境中比传统的TFIDF做得更好（只差几个百分点）。该方法还改进了Wieting等人的语句嵌入，如表1所示。最后，我们发现，与普遍的信念相反，Word2Vec（CBOW）在模型中也没有使用简单的词向量平均值，正如常用表达式所误导的那样。对实现的深入研究表明，它隐式地使用了词向量的加权平均值——同样，与TF-IDF不同——结果表明，这种加权在效果上与我们的非常相似。（见第3.1节）

**2 RELATED WORK**

单词嵌入。单词嵌入方法将单词表示为低维空间中的连续向量，该空间捕获单词的词汇和语义属性。它们可以从文本的神经网络模型（Bengio et al.，2003；Collobert&Weston，2008；Mikolov et al.，2013a）或共生统计的低阶近似（Deerwester et al.，1990；Pennington et al.，2014）的内部表示中获得。已知这两种方法密切相关（Levy&Goldberg，2014；Hashimoto等人，2016；Arora等人，2016）。

我们的工作是与（Arora等人，2016）最直接相关的工作，该工作提出了在文档中生成单词的随机游走模型。我们的句子向量可以看作是潜在变量在生成模型中的近似推理。

短语/句子/段落嵌入。以前的工作是通过对向量和矩阵的运算来组合单词嵌入来计算短语或句子嵌入，例如（Mitchell&Lapata，2008；2010；Blacoe&Lapata，2012）。他们发现，在所研究的二进制运算中，矢量的坐标相乘表现得非常好。未加权平均法也能很好地表示短语（Mikolov等人，2013a）。另一种方法是在解析树上定义的递归神经网络（RNN），在监督下训练（Socher et al.，2011）或不训练（Socher et al.，2014）。简单rnn可以看作是一种特殊情况，解析树被简单的线性链所代替。例如，skip gram模型（Mikolov等人，2013b）被扩展为包含序列的潜在向量，或者将序列而不是单词作为基本单元。在（Le&Mikolov，2014）中，每一段都被认为具有潜在的段落向量，它影响段落中单词的分布。Skip Think of（Kiros et al.，2015）试图从包围的句子中重建包围的句子，并将隐藏的参数视为它们的向量表示。使用长-短期记忆（LSTM）的RNN捕获长距离依赖性，也用于句子建模（Tai等人，2015）。

其他神经网络结构包括卷积神经网络，例如（Blunsom等人，2014年），它使用动态池处理不同长度的输入语句，并在情绪预测和分类任务中表现良好。

对我们工作的直接启示是（Wieting等人，2016年），通过使用简单的单词平均法学习释义句子嵌入，并基于释义对的监督更新标准单词嵌入；监督用于初始化和培训。

**3 A SIMPLE METHOD FOR SENTENCE EMBEDDING**

我们简要回顾了中文本的潜在变量生成模型（Arora等人，2016）。该模型把语料库的生成看作一个动态过程，其中第t个词是在步骤t产生的。这个过程是由话语向量的随机游走驱动的。词汇中的每个词w也有一个向量in ；这些都是模型的潜在变量。语篇向量表示“正在谈论的内容”。语篇向量ct和（时不变）词向量（对应词w）之间的内积捕获语篇和词之间的相关性。在时间t观察单词w的概率由Mnih和Hinton的对数线性单词产生模型给出：

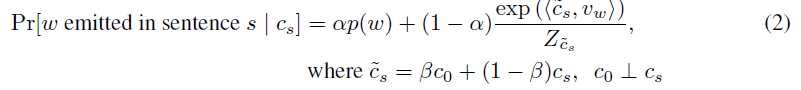


语篇向量ct做缓慢的随机游走（即ct+1是通过添加一个小的随机位移向量从ct中获得的），从而在相似的语篇下生成邻近的词。在（Arora等人，2016年）中表明，在一些合理的假设下，该模型会产生行为——根据符合word2vec和Glove等经验性著作的单词共现概率。随机游走模型可以放宽到允许ct中偶尔出现大的跳跃，因为一个简单的计算表明它们对词的共现概率的影响可以忽略不计。据报道，使用该模型计算的词向量与Glove和word2vec（CBOW）计算的词向量相似。

我们改进的随机游走模型。显然，将句子嵌入定义为：给定一个句子s，对支配这个句子的语篇向量做一个映射估计。我们注意到，我们假设当句子中的词被发出时，语篇向量ct没有太大的变化，因此我们可以用一个语篇向量cs替换句子中的所有ct。本文（Arora et al.，2016）表明，cs的MAP估计值等于句子中单词嵌入的平均值乘以标量。

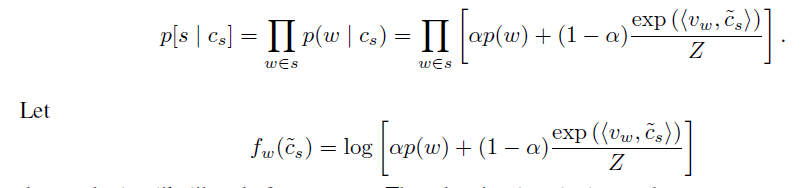
在本文中，为了更真实地建模，我们将模型（1）更改如下。该模型有两种类型的“平滑词”，这是为了解释这样一个事实，即有些词出现在语境之外，而有些频繁出现的词（可能是“the”、“and”等）则常常出现在不考虑话语的情况下。我们首先在对数线性模型中引入了一个加法项，其中p（w）是单词（在整个语料库中）的单格概率，而是标量。这允许单词出现，即使它们的向量与cs的内积很低。其次，我们引入了一个常用的语篇向量，它可以作为与句法相关的最频繁语篇的修正项。（其他可能的修正留待将来的工作）它提高了沿c0有高成分的词的共现概率。

具体地说，给定语篇向量cs，句子s中的单词w发出的概率由以下公式建模：，

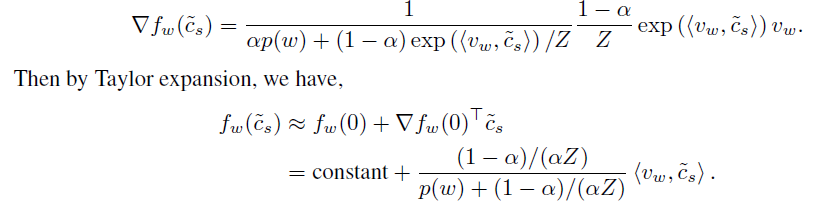


其中和是标量超参数，是规范化常数（分区函数）。我们看到这个模型允许一个与话语无关的词w来讨论cs的发射有两个原因：a）来自术语 p（w）；b）如果w与公共语篇向量c0相关。

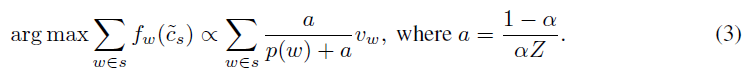
计算句子嵌入。我们的模型生成的单词嵌入实际上是相同的。句子嵌入将被定义为生成它的向量cs的最大似然估计。（在这种情况下，MLE与MAP相同，因为前者是一致的。）我们借用了（Arora et al.，2016）的关键建模假设，即单词vw的大致均匀分布，这意味着分区函数Zc在所有方向上大致相同。所以假设大致相同，比如Z代表所有的~cs。根据模型（2），判刑的可能性是



表示句子s的对数似然。然后，通过简单的演算我们得到，

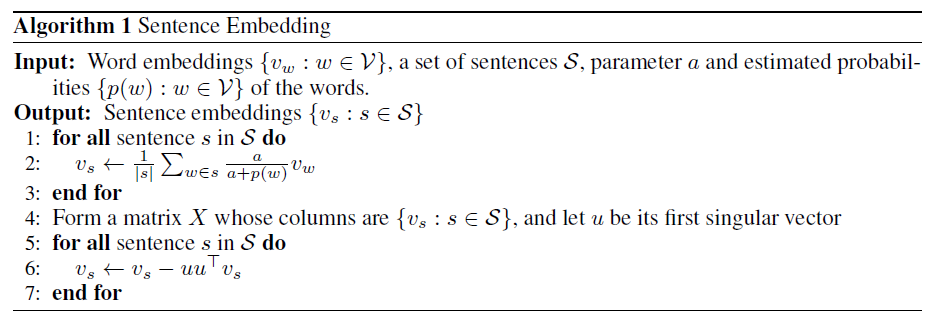


因此，单位球面上~cs的最大似然估计（忽略正规化）近似为



也就是说，MLE大约是句子中单词向量的加权平均值。注意，对于更频繁的单词w，权重a=（p（w）+a）较小，因此这自然会导致频繁单词的权重降低。

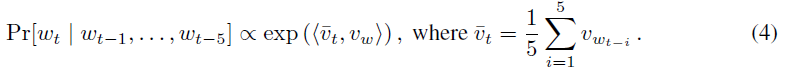
为了估计cs，我们通过计算一组句子的~cs's的第一主成分来估计方向c0。换句话说，最后的句子嵌入是通过减去~cs's对其第一主成分的投影得到的。算法1对此进行了总结。



3.1与WORD2VEC中的子抽样概率的关系

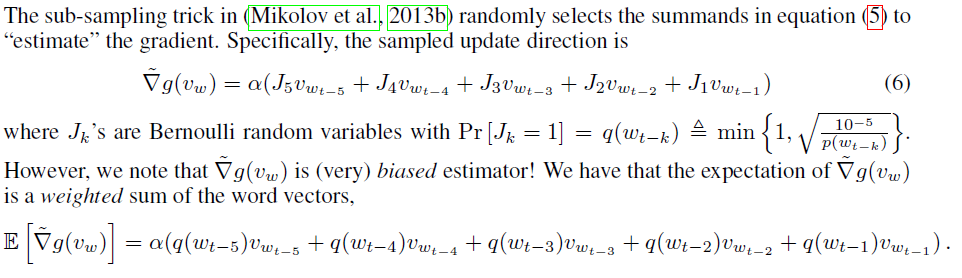
Word2vec（Mikolov等人，2013b）使用了一种子抽样技术，该技术以与成比例的概率对单词w进行下采样，其中p（w）是单词w的边际概率。这种启发式方法不仅加快了训练速度，而且还学习了更多的规则单词表示。在这里，我们解释说，这对应于模型中的词向量的隐式重新加权，因此统计上的好处应该毫不奇怪。

回想一下word2vec的普通CBOW模型：



可以证明，单字向量vw的损失（MLE）可以抽象地写成这种形式，





实际上，期望值对应于修正的word2vec模型的梯度，平均值vt（在等式（4）中）被加权平均所代替。这样的加权模型也可以与我们从随机游动模型中得到的公式（3）相同。此外，当使用参数时，加权q（wi）密切跟踪我们的加权方案a=（a+p（w））；见图1。因此，这里的期望梯度是我们模型中估计的语篇向量！因此，具有次采样梯度启发式的word2vec对应于使用我们的加权方案的随机梯度更新方法。

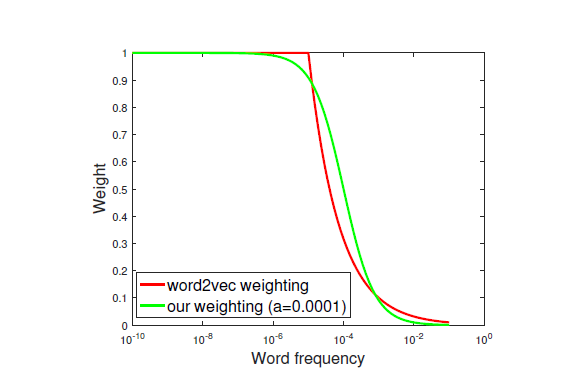


图1:word2vec中的子抽样概率与我们的加权方案相似。

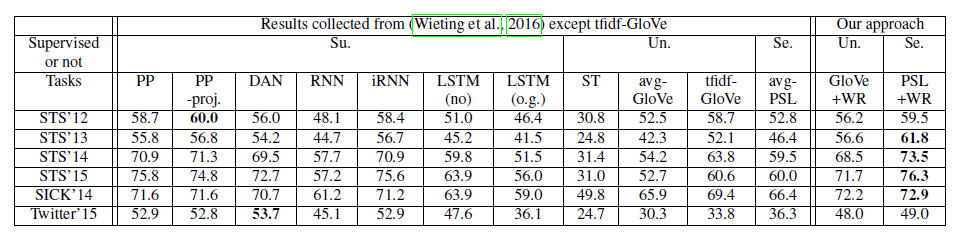


表1：文本相似性任务的实验结果（皮尔逊rX100）。每行的最高分是黑体字。这些方法可以被监督（表示为Su.）、半监督（Se.）或无监督（Un.）。“GloVe+WR”表示将我们的方法应用于GloVe词向量得到的句子嵌入；“PSL+WR”表示PSL词向量。有关方法的说明，请参见正文。

**4实验**

**4.1文本相似性任务**

数据集。我们在22个文本相似性数据集上测试了我们的方法，包括SemEval语义文本相似性（STS）任务（2012-2015）（Agirre等人，2012；2013；2014；Agirrea等人，2015）和SemEval 2015 Twitter任务（Xu等人，2015）和SemEval 2014语义相关性任务（Marelli等人，2014）的所有数据集。这些任务的目的是预测两个给定句子之间的相似性。评价标准是预测分数和地面真实分数之间的皮尔逊系数

实验设置。我们将把我们的方法与以下方法进行比较：

一。无监督：ST、avg手套、tfidf手套。ST表示跳跃思维向量（Kiros等人，2015），avg globe表示手套向量的未加权平均值（Pennington等人，2014），5和TF IDF globe表示使用TF-IDF权重的手套向量的加权平均值。

2。半监督：平均PSL。该方法使用段落999（PSL）词向量的未加权平均值（Wieting等人，2015）。使用标记数据训练词向量，而句子嵌入则是在不训练的情况下用未加权平均值计算的。

三。监督：PP、PP项目、DAN、RNN、iRNN、LSTM（o.g.）、LSTM（no）。所有这些方法都使用PSL词向量初始化，然后在PPDB数据集上进行训练。PP和PPproj。

建议于（Wieting等人，2016年）。第一个是词向量的平均值，第二个额外添加一个线性投影。单词向量在训练期间更新。DAN表示深度平均网络（Iyyer等人，2015）。RNN表示经典的递归神经网络，iRNN表示激活为恒等式的变量，权值矩阵初始化为恒等式。LSTM是（Gers等人，2002）的版本，有输出门（表示为LSTM（o.g.））或没有（表示为LSTM（no））。

我们的方法可以应用于任何类型的字嵌入。因此，我们将我们的方法应用于单词嵌入方法“XXX”得到的句子嵌入表示为“XXX+WR”。为了得到一个完全无监督的方法，我们将其应用于手套向量，表示为手套+WR。将加权参数a固定为10~3，单词频率p（w）由commoncrawl数据集。这在表1中用手套+WR表示。我们还将我们的方法应用于PSL向量，表示为PSL+WR，这是一种半监督方法。

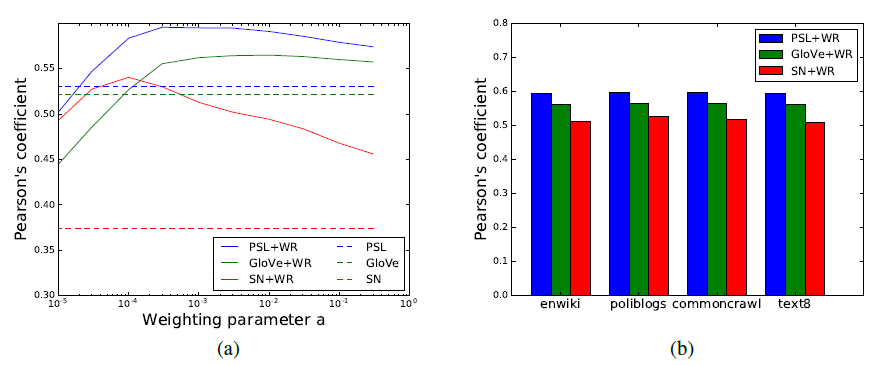


图2：我们方法中的加权方案对STS2012任务的平均性能的影响。最好用彩色观看。（a） 性能v.s.加权参数a.使用在enwiki数据集上估计的p（w）对三种类型的词向量（PSL、GloVe、SN）进行测试。最佳性能通常在a=10-3至a=10-4时达到。（b） 用于估计p（w）的性能v.s.数据集。四个数据集（enwiki，poliblogs，commoncrawl，text8）用来估计p（w），然后在我们的方法中使用。参数a固定为10-3。对于不同的设置，性能几乎相同。

结果。结果见表1。每年有4到6个STS任务。为了清楚起见，我们只报告每年STS任务的平均结果；详细结果见附录。

无监督方法手套+WR比avg手套显著提高了10%-30%，大大超过了基线。它比LSTM和RNN具有更好的性能，并且可以与DAN相媲美，即使后面的三个使用监控也是如此。这证明了这种简单方法的威力：它甚至比高度优化的、经过监督训练的复杂模型更强大。使用TF-IDF加权方案也比未加权平均值有所改善，但不如我们的方法。

半监督方法PSL+WR在六项任务中有四项取得了最好的结果，并且在其余两项任务中可与之相媲美。总的来说，它优于平均PSL基线和用相同PSL向量初始化的所有监督模型。这说明了我们的方法比那些模型的训练更具优势。

我们还注意到，数据集的顶部奇异向量c0似乎大致对应于句法信息或常用词。例如，在病态数据集中，与c0最接近的词（余弦相似）是“just”、“when”、“even”、“one”、“up”、“little”、“way”、“there”、“while”和“but”。最后，在附录中，我们展示了我们的两个想法都有助于改进：对于手套向量，单独使用平滑逆频率加权比未加权平均提高约5%，单独使用公共分量去除提高10%，使用两者提高13%。

4.1.1加权参数对性能的影响

我们研究了我们的方法对加权参数a、计算词向量的方法和估计的词概率p（w）的敏感性。首先，我们测试三个STS 2012任务中的词向量类型（PSL、GloVe和SN）。SN向量在enwiki数据集（Wikimedia，2012）上使用（Arora et al.，2016）中的方法进行训练，而PSL和手套向量在表1中使用。我们列举了一个，并使用了enwiki数据集上估计的p（w）。图2a显示，对于所有三种类型的词向量，较之未加权平均值，a的大范围导致显著改进的性能。最佳性能出现在a=10∏3到a=10∏4之间。

接下来，我们修正a=10а3，使用四个非常不同的数据集来估计p（w）：enwiki（维基百科，30亿代币）、poliblogs（Yano等人，2009年）（政治博客，500万）、commoncrawl（Buck等人，2014年）（互联网爬网，8000亿）、text8（Mahoney，2008年）（wiki子集，100万）。图2b显示了所有四种设置的性能几乎相同。

我们的方法可以应用于不同类型的词向量，在不同的语料库上进行训练，这也表明它应该在不同的领域中有用。这对于无监督方法尤其重要，因为可用的未标记数据可能收集在与目标应用程序不同的域中。

**4.2 SUPERVISED TASKS**

该方法得到的句子嵌入可以作为下游监督任务的特征。我们考虑了三个任务：病态相似性任务、病态蕴涵任务和斯坦福情绪树库（SST）二元分类任务（Socher等人，2013）。为了突出无监督学习句子嵌入的表现力，我们对嵌入进行了修正，只学习分类器。监督任务的设置主要遵循（Wieting等人，2016）以允许公平比较，即分类器a线性投影，然后是分类器（Kiros等人，2015）。线性投影将句子嵌入2400维（与跳过的思想向量相同），并在训练过程中学习。我们将我们的方法与第4.1节中使用的PP、DAN、RNN和LSTM进行比较。我们还比较了跳跃思维向量（在（Lei Ba等人，2016年）中改进的训练）。

结果。与竞争对手相比，我们的方法获得更好或可比的性能。对于其中两个任务，它获得了最好的结果。这证明了我们简单方法的威力。我们强调我们的嵌入是在无监督的情况下学习的，而DAN、RNN、LSTM则是在有监督的情况下训练的。此外，skip思想向量的维数比我们的要高得多（尽管投影到更高的维数中，原始的300维嵌入包含了所有的信息）。

在文本相似性任务中，这种优势没有那么显著。这可能是因为相似性任务直接依赖于余弦相似性，这有利于我们的方法去除公共成分（可以看作是去噪的一种形式），而在有监督的任务中，由于一些标签信息的代价，分类器可以挑选出有用的成分而忽略公共成分。

最后，我们推测我们的方法在情感任务上没有优于RNN和LSTM的方法，因为（a）单词向量-更广泛地说是意义的分布假设-由于“反义词问题”在捕捉情感上有已知的局限性，（b）在我们的加权平均方案中，像“not”这样的单词可能是对于情绪分析来说很重要的是被低估了很多。为了解决（a）的问题，现有的工作是学习更好的单词嵌入以进行情绪分析（例如，Maas等人，2011年）。为了解决（b），可以为这个特定任务设计权重方案（或学习权重）。

4.3句子语序的影响

我们方法的一个有趣的特点是它忽略了词序。与此相反，RNN和LSTM可以潜在地利用词序。我们的方法在这些基准上取得了更好或可比的性能，这一事实提出了以下问题：单词顺序在这些基准中不重要吗？我们做了一个实验表明语序确实起到了一定的作用。

我们在监督任务中对RNN/LSTM进行了训练和测试，其中每个句子中的单词被随机洗牌，结果如表3.所示，可以观察到性能明显下降。因此，我们忽略语序的方法必须比RNN和LSTM的方法更好地利用语义，一个有趣的未来方向是探索集成思想是否能将这两种方法的优点结合起来。

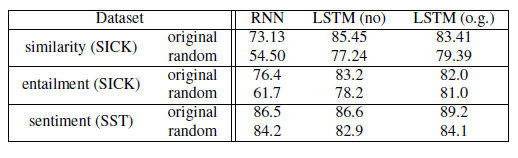


表3：原始数据集和句子中随机洗牌的结果的比较。标签为“原始”的行是原始数据集上的结果，标签为“随机”的行是随机洗牌数据集上的结果。相似性行（SICK）显示Pearson的rX100，其他行显示准确性。

5 CONCLUSIONS

这项工作提供了一种简单的句子嵌入方法，基于生成文本的随机游走模型中的话语向量（Arora等人，2016）。它简单、无监督，但在各种文本相似性任务上的性能明显优于基线，甚至可以胜过一些复杂的有监督方法，如RNN和LSTM模型。所得到的句子嵌入可以作为下游监督任务的特征，与成熟的方法相比，也可以得到更好或可比的结果。