实验报告

近期，我在进行自然语言处理领域的相关科研，但每次实验的目标都注重于模型的优化，一直用简单的word2vec工具生成词向量，未考虑过词向量的质量对实验效果的影响。借此机会，我**比较了七种词向量生成模型在各个自然语言处理领域的八项任务上的效果，并研究了词向量维度和语料库规模对词向量性能的影响。**
以下从词向量生成和词向量评估两个过程对此次实验进行阐述。

词向量生成

在复现论文的过程中，考虑到时间效率，我只选用了**100M的语料库**，运用**skip-gram模型**，进行**5次迭代**，生成**200维**词向量。因为skip-gram具有随机性，这个生成过程**执行三次**，并对生成的词向量进行评估。
该词向量生成过程语料库小，迭代次数少，但也运行了近6个小时，这说明生成一个足够优秀的词向量需要巨大的GPU资源，个人电脑无法支持如此巨大的计算量。

说明：所运用的语料库为embedding文件夹中的text8.txt文件，生成的词向量为vec.txt_1,vec.txt_2,vec.txt_3。

词向量评估

avg任务（文本分类任务）

实验任务为：
使用词向量的加权平均值作为文本的表示，然后应用逻辑回归来执行文本分类。 每个单词的权重是其出现频率。使用的数据集是IMDB数据集。

实验结果为：
Accuracy = **75.3031%** (18822/24995)
Accuracy = 75.001% (18748/24997)

论文中的词向量参数设置为28亿个单词短语的W&N的语料库和50维词向量，最好准确率为**74.94%**，而我的实验在该任务上的准确率超过了论文中最优准确率，这证明我的参数设置在该问题上能得到更优的词向量。当然，这一结论需要更多次实验进行论证。

cnn任务（句子情感分析）

实验任务为：
使用卷积神经网络(CNN)在斯坦福情绪树库数据集上进行句子情感分类，重复实验五次，并展示这些实验的平均准确性。

实验结果为：
train:0.067420(99.64%,99.59%), valid:2.207849(36.24%,34.38%), test:2.286111(35.48%,34.23%)
train:0.067051(99.63%,99.58%), valid:2.184892(**37.33%**,35.70%), test:2.247548(35.79%,34.57%)
train:0.077034(99.51%,99.47%), valid:2.329713(35.42%,33.26%), test:2.231513(36.20%,33.75%)
train:0.063223(99.64%,99.59%), valid:2.251042(33.33%,32.21%), test:2.232813(35.52%,34.58%)
train:0.062751(**99.68%**,99.64%), valid:2.220478(36.69%,34.70%), test:2.162657(**37.15%**,35.70%)

可以发现该任务在训练集上的准确率高于99%，但在验证集和测试集上的准确率明显偏低，证明存在过拟合的现象。论文中的最优准确率为**43.84%**，我的结果和论文实验结果仍有差距。

pos任务（词性标注）

实验任务为：
使用Ronan Collobert及其同事提出的神经网络对华尔街日报数据进行词性标注，并评估准确性。

实验结果为：
train:0.273759(92.01%), valid:0.540134(85.50%), test:0.522921(**85.89%**)

论文中最优准确率为**96.57%**，而我的准确率为85.89%，这说明该任务对于语料库的大小要求比较高，需要较高的语料库提高泛化能力以提高准确率。

syn-sem任务（语义类比与语法类比）

实验任务为：
完成大约9000个语义类比和语法类比问题，问题类似于“ man is to (woman) as king is to queen”。通过计算（queen-king+man）的最近词向量作为问题的答案，并评估整体的准确性。

实验结果为：
max total accuracy: 12.45%
max semantic accuracy: **12.45%**
max syntactic accuracy: **12.08%**

而论文中的语义类比最优准确率为**51.78%**，语法类比最优准确率为**44.80%**，可见与词性标注相比，语义类比和语法类比对于语料库的规模有更高的要求，过小的语料库会导致实验准确率过低。

tfl任务（同义词选择题）

实验任务为：
测试过程用的是托福考试中的80个同义词选择题，每道选择题有四个选项，选择问题与选项中余弦距离最近的选项，并评估最终整体的准确性。

实验结果为：
accuracy： **48.75%**

论文中最优实验结果为**76.25%**，说明同义词选择题对于语料库的规模也有一定要求。

ws任务（词语相似度）

实验任务为：
测试过程用的是WordSimilarity-353测试集，他包含353对英文词汇和人工对这些词

对之间的语义相关度的评测值。词向量的效果比较的是两个词向量的余弦距离的Pearson相关性与人工打分的平均分数的相似程度。

实验结果为：
accuracy： **60.52%**

论文中的最优准确率为**63.89%**，我的实验结果略低于实验最优结果。

实验结论

某些任务对语料库的规模要求比较严格，其中syn-sem任务的表现最为明显；某些任务对语料库的规模要求较小，即使语料库较小也不会太影响准确率，换句话说，语料库规模的增加对该任务的准确率没有太大效果，例如ws任务、cnn任务等。在这次实验中，比较特殊的情况是，avg任务我的实验准确率优于论文中的最高准确率，这证明对于这个任务而言，词向量的维度比语料库的规模更重要，当然这一结论需要更多的实验支持。
遗憾的是，这个实验对于运算能力的要求比较高，作者提到，其实验花费了10万CPU小时，受到实验资源的限制，很多模型的比较我难以进行，但是我掌握了整个比较的方法和思维，对于我在自然语言处理领域的研究也大有裨益。