Word2Vec 及 BERT 模型的水稻文献词汇嵌入计算

邓启东 [1]

[1] 华中农业大学信息学院，430070，武汉，湖北，中国

摘要

水稻由于其营养价值高，加工副产品用途广作为主要的粮食作物。有关水稻文献文本挖掘方面的研究十分重要。

本次课程论文在 Pubmed 文献数据库中以 rice 作为关键词获取 6,859 篇水稻文献摘要文本，并且利用经典的词嵌入方法 Word2Vec 的得到其中高频词的嵌入向量并进行 t-SNE 可视化以直观呈现水稻文献词汇间的相似性关系。并与 BERT 等方法进行比较。最终凸显了 BERT 或 Bio-BERT 预先提供学习好的嵌入所带来的意义。

关键词：水稻，词嵌入，t-SNE，Word2Vec

1 课题概况

选修课程的目的主要在于我们生信的培养方案中涉及神经网络深度学习的内容较少。在学习了 perl、R、python、shell 等多种编程语言之后。而自然语言处理这门课程和生物信息学相结合，从生物文本中抽取知识，是学习一些比较智能的算法学习机会。在前面几次作业当中，我们发挥了充分的想象力进行下游分析。运用基础的代码进行了自主的探索。

选择嵌入计算的原因也在于 Word2Vec 算法本身设计的精妙。词语的向量分布式表示和神经网络的权重矩阵不谋而合。通过梯度下降逐渐找到一套合适的嵌入满足整个上下文文本。而且基于的是词语上下文相似则其语义相似的假设。相比于以往的各种方法而言是一大飞跃（在后文中也有所介绍）。可以使用相似性计算，取交叉熵的方式反映词汇之间的相似性。

本次课程论文计划能够很好的求出水稻文献中词汇的嵌入，并且通过 t-SNE 进行降维可视化。能够直观的感受到距离相近的词语的语义存在某种相似关联。就说明我们的算法是成功的。

2 数据

数据详见 reference.table.txt，其搜集方式是通过在 Pubmed 上以 rice 作为关键词检索下载得到。

除去标题行共计 6,859 篇摘要，并且分为六列：标题、年份、期刊、发文机构、和基因（以 “|” 分隔）。

我们将其中的摘要抽取出来，将其中的特殊标点符号!”♯$%’()*+,./:;<=>?@[\]^_‘{|}~ 全部替换成空格；并且将多个空格替换成一个空格，然后将单词全部转换成小写。把多个数字或者数字 + 字符 + 数字的组合替换成 NBR。

至此，我们得到了以空格分开的每个单词构成的语料库，保存在 data/corpus.txt 路径下。语料库数据准备完毕。

3 研究方法

3.1 研究方法的算法背景，与其他方法的联系与区别

在自然语言处理里面，最小的单位量是词语，词语组成句子，句子组成段落，段落在组成一篇文章，所以在处理自然语言的问题时，首先要对最基本的词语进行处理。

3.1.1 词典

我们想要表示一个词语，首先想到的是建立一个词典，而已经有这样一个词库 WordNet[1] 根据不同的词性建立起词和词的关系。实现词语分类，它是一种离散表征，反应不出词汇之间的差别。

这种方式存在以下两个弊端：

1. 缺少新词的含义；
2. 并不能完全的整合所有词，即便可以，它的数量也是十分庞大的。

由于词语是符号形式的，计算机也无法理解，因此需要转换成数值形式。

3.1.2 独热编码

使用独热编码（one-hot）进行表示。One-Hot 在特征提取上属于词袋模型（bag of words）。

独热编码有明显的缺陷：

1. 它不能表示所有的词, 即从未出现过的合成词语
2. 无法表示和学习词语之间的相关性，反映不出文字天然的内在含义，每个词都是正交的，即所有的词语点积均为 0，找不出相似词语
3. 会产生数据稀疏的问题，会浪费很多的存储空间解决它的方法是利用派生词法 (derivational morphology)，也就是使用词根词缀的方式来避免一味增加词典的问题。但是派生词法本身也存在致命的问题，就是有的词根意义实在太多，重复的意思同样也会伤害词的表意。

3.1.3 矩阵分解

通过 SVD 或者 PCA 降维等方式，将稀疏矩阵进行浓缩，得到一个低纬度稠密的类似矩阵。能更好的精炼词向量，并减少运算量。但是，这样的方法太过暴力，所得到的词向量效果也不好。

3.1.4 语义分布表示 (distributed representation)

最早由 Hinton 提出，可以克服 one-hot representation 的上述缺点，基本思路是通过训练将每个词映射成一个固定长度的短向量，所有这些向量就构成一个词向量空间，每一个向量可视为该空间上的一个点。此时向量长度可以自由选择，与词典规模无关。这是非常大的优势。

自然语言处理最基本的单位就是词语。但是词语本身比如中文、英文都是符号形式的，如果想要构建数学模型，必须要转化为数值型的输入。或者说——嵌入到一个数学空间里，这种嵌入方式，就叫词嵌入（word embedding)，而 Word2vec，就是词嵌入（word embedding) 的一种。

Word2Vec 是一种有效创建词嵌入的方法，它是从大量文本预料中以无监督方式学习语义知识的模型，这个模型为浅层双层的神经网络，用来训练以重新建构语言学之词文本。

Word2Vec 是轻量级的神经网络，其模型仅仅包括输入层，隐藏层和输出层，模型框架根据输入输出的不同，可以分为 CBOW 模型和 skip-gram 模型，CBOW 模型是通过上下文的内容预测中心词的可能情况，而 skip-gram 模型与其相反，它是通过中心词预测上下文词 [2]。

3.2 研究方法中的核心思路

Skip-gram 进行上下文预测的算法重点在于不断滑动改变中心词获得上下文最终生成批处理数据的过程。在这里我们使用“图 *1*”进行一个详细的展示，一些文献也对 Skip-gram 中的负采样有详细的描述 [3]。

首先黑色框的部分是我们的语料库，并且由于我们已经有了一个独热编码的词语字典。因此语料库中的每一个词都可以转变为一个 index。

下面蓝色的部分是一个 buffer，它就像是一个纸带一样从整个语料库的左端向有段滑动。这里规定 skip_window（跳跃窗口）为 4，也就是每个中心词单侧的词汇量。左右两边再加上中心词自身就是一个 buffer 的大小了, 因而这里为 9。skip_num 是每次跳跃窗口选取训练模型词语，这里我们设置为 8. 实际上它一般小于两倍 skip_window，我们这里取等于，意味着中心词上下的 8 个词语全部抽出。并且储存到 batch 里。

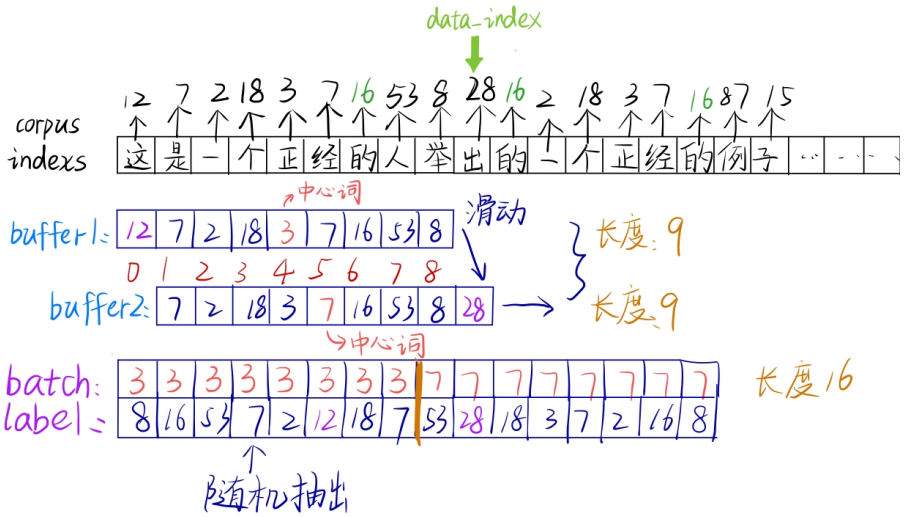


图 1: Skip-gram 算法介绍。图片来源：自绘　工具：GoodNotes

batch 的 3 和 7 是中心词的 index，下面的标签是抽出的上下文词汇。我们这里绘制的 batch_size 是 16，但是实际上在代码中是 128。包括学习率在内，这些超参数均可以修改，而且会影响代码运行的时间和效果。

3.3 本文的方法部分与课堂讲授内容的联系和区别/或补充

本文其实和第七次实验最大的区别在于。我们拿到的语料库是直接的文本，需要自己先对其使用 nltk 包进行预处理。

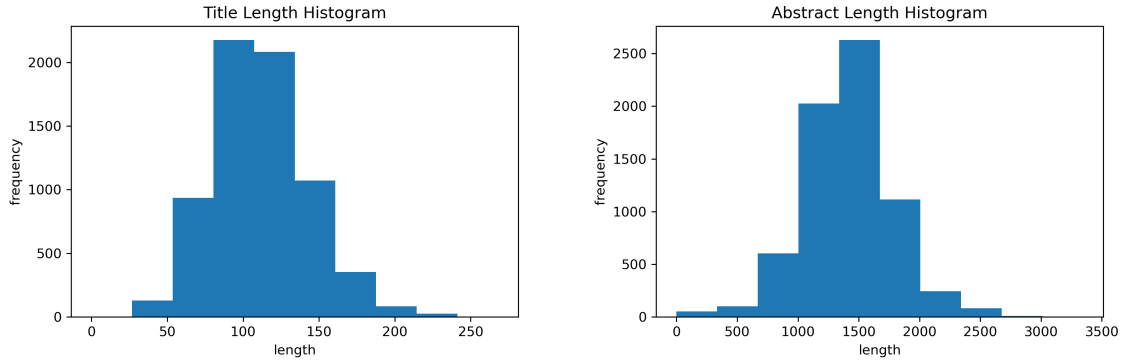
3.4 任务描述

请详述本任务的主要设计，并解释所设计算法和代码能适应于此任务的原因。

3.5 实验设计

3.5.1 语料库观察分析

我们首先对语料库的 title 和 abstract 的字符数目进行一个统计（图2）。其中 (a) 和（b）分别是这 6,859 篇摘要的标题和正文的长度频率分布直方图。其实整体可以看到存在为空的，尤其是 abstract。但是最终我们都会把他们一起合并成语料库而且去掉其标点。所以去不去掉空的行对实验没有影响。

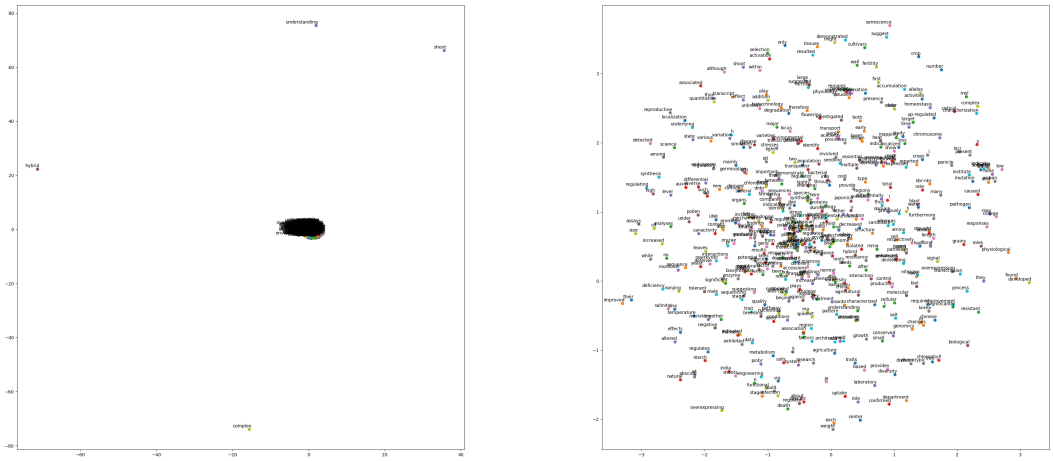


(a) 标题字符长度统计　　(b) 摘要字符长度统计

图 2: 语料库字符长度统计

3.6 某些关键代码

```
代码来源：Tutorial_4_word2vec-main/Skip_Gram_basic.py第六步绘图
# Step 6: Visualize the embeddings.
def plot_with_labels(low_dim_embs, labels, filename='tsne.png'):
  assert low_dim_embs.shape[0] >= len(labels), "More labels than embeddings"
  print('Visualizing.')
  plt.figure(figsize=(18, 18))  #in inches
  for i, label in enumerate(labels):
    x, y = low_dim_embs[i,:]
    #if x< -25 or x>35 or y < -25 or y > 25:
        #continue
    plt.scatter(x, y)
    plt.annotate(label,
                 xy=(x, y),
                 xytext=(5, 2),
                 textcoords='offset points',
                 ha='right',
                 va='bottom')
  plt.savefig(filename)
  print('TSNE visualization is completed, saved in {0}.'.format(filename))
```



(a) 少数单词过滤前　　(b) 控制坐标后

图 3: t-SNE 可视化结果

分析这张图片 (图3(a)) 的坐标轴尺度为什么不太合适。我们看到原因是存在四个词语的嵌入非常的偏离大部分词语，他们分别是 understanding、hybrid、shoot、complex，甚至达到了-70。原因是他们的上下文可能比较没有规律或者词频比较低。而其他的词语都在 3 上下。

比如当中绿色注释的两行通过限定 x,y 的范围都控制在 25 以内，如果超出 countinue 不参与绘图。以这样的方式过滤掉了比较偏离的那四个点。见图3(b),t-SNE 绘图的坐标轴变得合理了。

4　主要的生物信息学实验和实验结论

4.1　结果分析

我们从图中局部看到一些关联性。比如 root 和 seed 的距离就很近

此外还有 salinity 和 temperature 等环境因素。

本次实验的关联似乎没有第七次实验好，原因是训练的 num_steps 只有 1000，之后我们将其增加到 8000，得到新的图片，不过区别不大，命名为 tsne1.png

5　后记

2021 年 5 月 22 日 13 时 07，共和国勋章”获得者、中国工程院院士袁隆平老先生永远离开了我们。水稻作为我们亚洲的人口的主粮，袁隆平老先生培育的更加优秀的籼型杂交水稻对消除全世界人民的饥饿和贫困功不可没。就是这看似简单的一口饭，在过去的困苦年代是很难想象的。

见微知著，更加从这一次小的实验中体现出了 BioNLP，或者说科研的意义。其实早在实体识别那一次作业就有涉及到水稻 pubtator 的实体识别，并且也发了一些有趣的联系。

6　github

github 链接：https://github.com/LianzePuppet/article

参考文献

[1] D. Lin. Review of: Wordnet: An electronic lexical database. *computational linguistics*, 2002.

[2] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *Computer Science*, 2013.

[3] Y. Goldberg and O. Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv*, 2014.