学　　期 2021-2022（2）

北京航空航天大學
BEIHANG UNIVERSITY

深度学习与自然语言处理作业

词向量的应用

院（系）名称　自动化科学与电气工程学院

学　生　姓　名　黄旭聪

学　　号　ZY2103203

2022 年 5 月 20 日

1. 问题阐述

问题：利用给定语料库（或者自选语料库），利用神经语言模型（如：Word2Vec，GloVe 等模型）来训练词向量，通过对词向量的聚类或者其他方法来验证词向量的有效性。

词向量是将字词映射到向量空间的模型，目的是把字词转换成向量，从而能让计算机进行操作和运算。词向量也叫做 Embedding，可以看作是一种降维。在之前的作业中，我们主要学习了基于频率的词嵌入，例如 LDA 和 TF-IDF。这次大作业主要是基于神经网络，即基于预测的词嵌入。

2. 背景原理

2.1 One-hot 编码

One-hot编码是最简单的一种词向量方式，其用很长的向量表示一个词，向量的长度为词典的长度。即一个向量只有一个1，其他全为0，这样的稀疏性对于很大的词典来说，难以计算。且容易受维度灾难的困扰，尤其是用于深度学习的一些算法时。

2.2 分布式的表示

通过训练将每一个词映射成一个固定长度的短向量。所有这些向量构成一个词向量空间，而每一个向量则可视为该空间中的一个点。根据词之间的距离来判断它们之间的语法、语义上的相似性，这样就将词的信息分布到各个分量中了。例如上一次大作业的LDA语言模型就属于这类方法。

2.3 word2vec

本次大作业采用word2vec模型，word2vec是一种浅层的神经网络，由嵌入层、隐藏层、输出层构成。其根据输入，输出的特点可以分为两种模式：CBOW（由上下文预测当前词），Skip Fram（由当前词预测上下文）。如下图所示。这两种方法的优化目标都是在已知先验知识的基础上，使得预测目标值的极大似然估计值最大。为了估计极大似然估计值，即需要计算目标词汇出现的概率，该概率的计算需要涉及词汇表中的所有词汇。因此每次网络更新时，每一次预测都是基于全部的数据集进行的，时间开销很大。基于此，提出了两种加快训练速度的方法，一种是负采样，一种是二叉树式的层级结构。

他们的共同点都是将需要基于全部样本的计算变为基于部分值的计算。

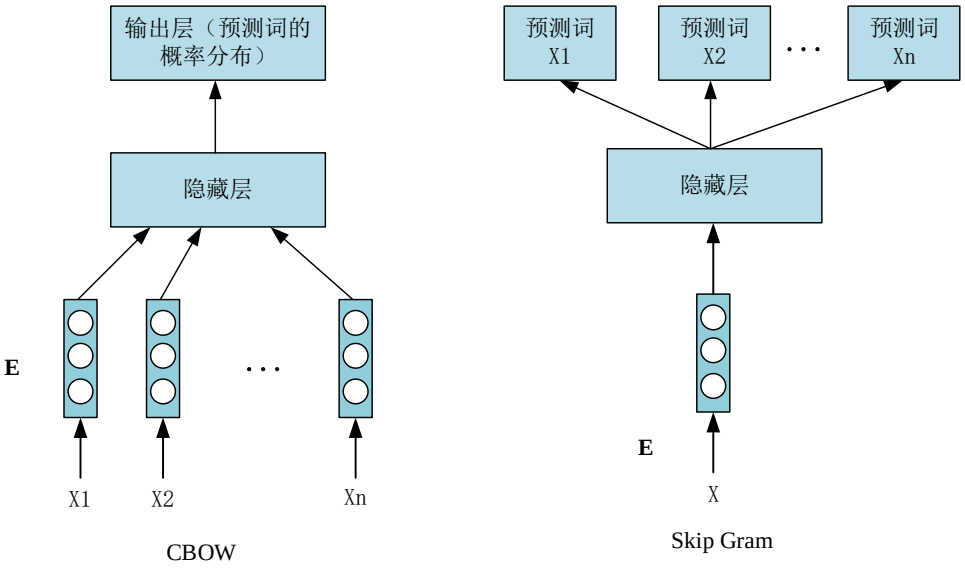

图 1 word2vec 两种网络结构

3. 实验部分

本文的训练语料为金庸的全部16本小说。具体分为数据预处理，制作corpus，训练得到词向量模型，相关分析。

3.1 数据预处理

在此前作业的基础上进行，对于分词后的结果进行比较，出现以下几种情况：1.一些无实际意义的词语频繁出现，例如“说道”，“各位”。2.由于分词是广义使用的，分词不够准确，例如“郭靖道”，“郭靖提”。针对这些问题，改进预处理过程。具体的，增加停用词的使用，在分词前提出无意义的词；此外，为了后续研究人物，武功，门派等关联，在jieba的词汇表中加入金庸小说中出现的人物名称、武功和门派。相关内容详见目录下的trainset文件夹。

3.2 制作corpus

由于本次大作业使用gensim库提供的word2vec接口，故需要提供自定义的嵌套存储列表。每一个句子为列表中的一个元素，每一个句子由由分好的词构成一个列表。总共生成16部小说下的208912句语料。

3.3 训练词向量模型

使用gensim库提供的word2vec。分别使用CBOW及Skip Gram模型，设定词向量的维数为200，低频词的过滤阈值为5，训练词与上下文词范围距离为5.分别得到训练好的CBOW及Skip Gram模型：model_cbow.model，model_skip.model。

3.4 相关分析

首先直观展示在金庸小说集中由意义的词对应的词向量，只展示词向量中最相关的上下文词。选用指定词包括：杨过,郭靖,段誉,降龙十八掌,全真剑法,武当派,屠龙刀,功夫。分别验证金庸小说中核心的人物，功夫，武器和门派。结果如下图展示。

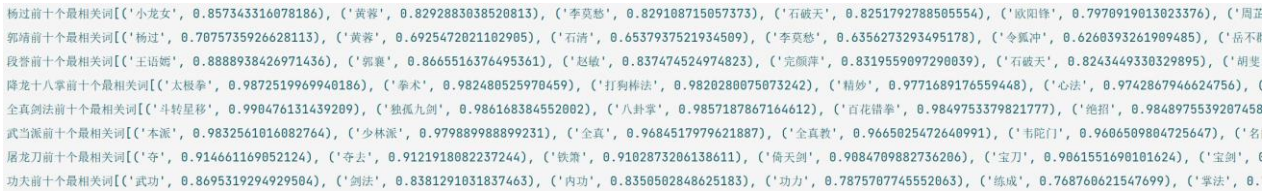

图 2 CBOW 核心词对应主要相关词

图 3 SkipGram 核心词对应主要相关词

可以发现，“杨过”最相关的词为“小龙女”，“郭靖”为“黄蓉”，“段誉”为“王语嫣”，“降龙十八掌为“打狗棒法”，“功夫”为“轻功”，“屠龙刀”为“倚天剑”等，这些对应关系与小说一致。

其次，验证相关词的推理功能：

```
----------------------------------------
相关词推理
根据郭靖+黄蓉=小龙女+X,找到X:杨过,对应概率：0.5278612375259399
根据黄蓉+打狗棒法=洪七公+X，找到X:招中,对应概率：0.7733055949211121
郭靖\小龙女\门户\段誉中排除异类的词是:门户
乔峰\虚竹\段延庆\段誉中排除异类的词是:段延庆
阿朱\木婉清\王语嫣\阿碧中排除异类的词是:阿碧
----------------------------------------
```

图 4 CBOW 相关词推理

```
----------------------------------------
相关词推理
根据郭靖+黄蓉=小龙女+X,找到X:洪七公,对应概率: 0.46012258529663086
根据黄蓉+打狗棒法=洪七公+X, 找到X:棒法,对应概率: 0.5812072157859802
郭靖\小龙女\门户\段誉中排除异类的词是:段誉
乔峰\虚竹\段延庆\段誉中排除异类的词是:乔峰
阿朱\木婉清\王语嫣\阿碧中排除异类的词是:木婉清
----------------------------------------
```

图 5 SkipGram 相关词推理

对比发现，可以发现运用 CBOW 学习到的词向量更准确，得出的相关词推理更符合逻辑。

最后，利用词向量的相似度进行一些探索。在《天龙八部》中，段誉与多名优秀的女英雄想认识。最初差点与木婉清成为一家人，到后来又认识了王语嫣，钟灵。根据文本情况，挖掘词向量的相似度。相似度可以一定程度反应共同进行的事情，相似度越高，说明词在文章中的出现较为靠近，两人所作的事情也更多，更加契合。对比结果如下图：

```
----------------------------------------
相似度
段誉喜欢王语嫣/钟灵/木婉清的程度:0.8888938426971436 0.7747688293457031 0.7789785861968994
----------------------------------------
```

图 6 CBOW 相似度的应用

```
----------------------------------------
相似度
段誉喜欢王语嫣/钟灵/木婉清的程度:0.7322624921798706 0.6345649361610413 0.6765795946121216
```

图 7 Skip Gram 相似度的应用

可以发现与文章相统一。

4. 总结

从结果看来，CBOW模型更注重文章整体的关联，很多相似度高的词跨越了多本小说；而Skip Gram模型更加注重局部。这与两者的使用范围相一致。CBOW适用于语料库较大时，Skip Gram适用于小语料库。

本次大作业对于词向量的构造以及应用有了进一步的了解。为后续基于深度学习相关方法的学习打下了基础。