NLP_REPORT_3

基于 Word2Vec 语言模型的词向量有效性验证

ZY2343226　朱子航

1010951286@qq.com

Abstract

基于所给的中文语料库文件（jyxstxtqj_downcc.com），利用 1～2 种神经语言模型（如：基于 Word2Vec ， LSTM， GloVe 等模型）来训练词向量，通过计算词向量之间的语意距离、某一类词语的聚类、某些段落直接的语意关联、或者其他方法来验证词向量的有效性。

Introduction

1.词向量

自然语言处理相关任务中要将自然语言交给机器学习中的算法来处理，通常需要将语言数学化，因为机器不是人，机器只认数学符号。向量是人把自然界的东西抽象出来交给机器处理的东西，可以说向量是人对机器输入的主要方式。

而词向量就是用来将语言中的词进行数学化的一种方式，也被称为词嵌入。它将单词映射到一个连续向量空间中的向量，使得具有相似含义的单词在向量空间中距离较近。词在送到神经网络训练之前需要将其编码成数值变量，常见的编码方式有两种：One-Hot Representation 和 Distributed Representation。

2.Word2Vec 模型

Word2vec，是一群用来产生词向量的相关模型。这些模型为浅而双层的神经网络，用来训练以重新建构语言学之词文本。网络以词表现，并且需猜测相邻位置的输入词，在 word2vec 中词袋模型假设下，词的顺序是不重要的。训练完成之后，word2vec 模型可用来映射每个词到一个向量，可用来表示词对词之间的关系，该向量为神经网络之隐藏层。

Word2Vec 主要包括 CBOW 模型（连续词袋模型）和 Skip-gram 模型（跳字模型。

2.1 .Skip-Gram 模型

Skip-Gram 模型的目标是根据中心词预测上下文词汇。具体而言，对于给定的中心词，模型的目标是最大化预测该中心词附近的上下文词的概率。在训练中，Skip-Gram 模型通过最大化给定中心词下上下文词出现的条件概率来调整词向量，以便使得这些词在向量空间中更接近。通过反向传播算法和梯度下降等优化方法，模型逐步调整词向量以最大化这些条件概率，从而学习到了每个单词的向量表示。

2.2 .连续词袋（CBOW）模型

CBOW 模型正好相反，它的目标是根据上下文词来预测中心词。具体而言，对于给定的上下文词，模型试图预测中心词的概率。CBOW 模型也是通过最大化给定上下文词的条件概率来训练词向量。模型通过调整词向量，使得该词在给定上下文的条件下具有最高的预测概率。类似于 Skip-Gram 模型，CBOW 模型也使用反向传播和梯度下降等方法来优化词向量。

Methodology

本文首先对语料库中数据进行处理，然后使用开源的 Gensim 库提供的接口来训练 Word2vec 模型，模型训练完毕后选择几本小说中的代表性人物或门派，分析训练后与该人物或门派相关性最强的几个词，来验证模型的有效性。

本文主要使用的模型为 CBOW 模型，给定一个长度为 T 的文本序列，设时间步的词为 W(t),背景窗口大小为 m，则连续词袋模型的目标函数（损失函数）是由背景词生成任一中心词的概率。

$$\sum_{t=1}^{T} P\left(w^{(t)} | w^{(t-m)}, \ldots, w^{(t-1)}, w^{(t+1)}, \ldots, w^{(t+m)}\right)$$

1.语料处理

在读取语料后，首先利用 jieba 分词对语料进行分词，去掉 txt 文本中一些无意义的广告和标点符号等内容，并将处理后的语料重新保存进新的文件夹中。

在实验过程中，注意到某些无关词语会在结果中出现，因此在进行语料处理时，去掉一些无关词语（例如，“我”、“你”、“他”、“她”、“它”等）

2.模型训练

使用开源的 Gensim 库提供的接口来训练 Word2vec 模型，调用的函数如下：

Word2Vec(sentences=LineSentence(name), hs=1, min_count=10, window=5,

vector_size=200, sg=0, epochs=200)

其中的参数的含义为：

- sentences：可以是一个 list，而对于大语料集，建议使用 BrownCorpus,Text8Corpus 或 LineSentence 构建；
- hs: 如果为 1 则会采用 hierarchical softmax 技巧。如果设置为 0（defaut），则 negative sampling 会被使用；
- min_count: 可以对字典做截断，词频少于 min_count 次数的单词会被丢弃掉, 默认值为 5，这里的值为 10；
- window：表示当前词与预测词在一个句子中的最大距离是多少，这里为 5；
- vector_size：是指特征向量的维度，默认为 100，这里使用 200；
- sg： 用于设置训练算法，默认为 0，对应 CBOW 算法；sg=1 则采用 skip-gram 算法；
- epochs：迭代次数，默认为 5，这里使用 200。

3.聚类分析

模型训练完毕之后，通过 Gensim 库中给出的接口函数，输出训练之后的模型，与某个给定输入词关联度最高的词或者是给定的某两个词之间的关联性。

本文选取了语料库中的《倚天屠龙记》、《天龙八部》、《射雕英雄传》、《神雕侠侣》、《笑傲江湖》等五本小说作为样本，对其中的代表性人物和门派分别进行了聚类分析。

Experimental Studies

将聚类的结果汇总为表格，整理如下：

表 1 关联度分析（倚天屠龙记）

倚天屠龙记			
张无忌关联度		明教关联度	
周芷若	0.515420377	本教	0.39372763
谢逊	0.458591789	魔教	0.336221009
张翠山	0.427928358	教	0.266606271
赵敏	0.419667184	僧侣	0.257586658
鹿杖客	0.409294903	峨嵋派	0.234774962
金花婆婆	0.385673463	阳	0.224332601

张无忌跟周芷若是青梅竹马，父亲为张翠山，义父为谢逊，后来跟赵敏在一起，金花婆婆是明教的紫衫龙王，也是小昭的母亲，均与张无忌相关。

明教常自称本教，也常被六大门派叫做魔教。

表 2 关联度分析（天龙八部）

天龙八部			
乔峰关联度		逍遥派关联度	
游坦之	0.37418744	童姥	0.268427819
全冠清	0.322064221	本门	0.245288432
萧峰	0.308975041	虚竹	0.236380547
赵钱孙	0.308591306	意中人	0.230974942
虚竹	0.302385271	苏星河	0.23094368
段誉	0.286942631	别派	0.2290999

游坦之是乔峰的劲敌，一生中与其难舍难分，全冠清是乔峰在丐帮的得力干部，萧峰是乔峰的后用名，虚竹和段誉是乔峰的结拜兄弟。

天山童姥是逍遥派前任掌门逍遥子的大弟子，逍遥派的首徒是苏星河，虚竹后来加入了逍遥派。

表 3 关联度分析（射雕英雄传）

射雕英雄传			
郭靖关联度		丐帮关联度	
黄蓉	0.603085756	白	0.262123197
欧阳克	0.491362393	轩辕台	0.249859944
洪七公	0.447010666	帮中	0.242160141
裘千仞	0.442116797	鲁有脚	0.238190755
完颜康	0.437415749	降龙十八掌	0.23616147
黄药师	0.369954079	竹棒	0.203803077

郭靖的老婆是黄蓉，师傅是洪七公，跟上述的人都有交集。

与丐帮相关的都是一些门内弟子和门派武功之类。

表 4 关联度分析（神雕侠侣）

神雕侠侣			
杨过关联度		全真派关联度	
小龙女	0.63205409	尹志平	0.302774876
周伯通	0.503448963	丘道长	0.294409037
赵志敬	0.446306974	石墓	0.265421331
绿萼	0.436547101	武功	0.261610657
李莫愁	0.430926919	全真	0.260315657
陆无双	0.416495025	南下	0.251286507

小龙女是杨过的姑姑，周伯通则是杨过的亦师亦友，李莫愁则和小龙女有些瓜葛，陆无双曾与杨过萍水相逢。

全真教相关的都是全真教里面相关的人以及跟全真教相关的门派。

表 5 关联度分析（笑傲江湖）

笑傲江湖			
令狐冲关联度		华山派关联度	
岳不群	0.541049063	本派	0.381254613
林平之	0.503553092	华山	0.377988279
仪琳	0.443270981	本门	0.332691669
盈盈	0.427830547	恒山	0.280264378
岳夫人	0.422909826	气功	0.275444031
岳灵珊	0.421629041	青城派	0.265014142

与令狐冲相关的是他的师傅、师弟、师妹等人，其余人都是与令狐冲有交集的人。

华山派相关的都是跟华山派相关的门派和地方。

本次实验针对语料库中的五本小说《倚天屠龙记》、《天龙八部》、《射雕英雄传》、《神雕侠侣》、《笑傲江湖》作为样本，对其中的代表性人物和门派分别进行了聚类分析，实验结果如上所示，其中人物与门派的关联度均较高，符合小说的实际情况，说明模型训练效果较好，可以验证词向量的有效性。

References

[1] https://blog.csdn.net/weixin_44966965/article/details/124732760