深度学习与自然语言处理第三次作业报告

吴宣余

ZY2303121

wxy7334@buaa.edu.cn

一、问题描述

利用给定语料库，利用1~2种神经语言模型(如：Word2Vec， LSTM，GloVe等模型)来训练词向量，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。

二、词向量

传统的自然语言处理将词看作是一个个孤立的符号，这样的处理方式对于系统处理不同的词语没有提供有用的信息。词映射(word embedding)实现了将一个不可量化的单词映射到一个实数向量。Word embedding能够表示出文档中单词的语义和与其他单词的相似性等关系。它已经被广泛应用在了推荐系统和文本分类中。Word2Vec模型则是Word embedding中广泛应用的模型。Word2Vec使用一层神经网络将one-hot（独热编码）形式的词向量映射到分布式形式的词向量。使用了Hierarchical softmax，negative sampling等技巧进行训练速度上的优化。

1 词的One-Hot表示

最简单的词表示方法是One-hot Representation，这种方法把每个词表示为一个很长的向量。向量的长度为词典的大小，向量中只有一个1，其他为0，1的位置对应词在词典中的位置。举例：你是谁，转换为One-Hot编码为，你[0 0 1]，是[0 1 0]，谁[1 0 0]。这种One-Hot编码如果采用稀疏方式存储，会是非常的简洁：也就是给每个词分配一个数字ID。但这种表示方法有几个缺点：

1. 容易受维度灾难的困扰，当词数量达到1千万的时候，词向量的大小变成了1千万维，假设使用一个bit来表示每一维，那么仅一个单词大概就需要0.12GB的内存；
2. 任意两个词之间都是孤立的，无法表示语义层面上词汇之间的相关信息；
3. 强稀疏性，只有一位是1，而其他位都是0，这就导致向量中有效的信息非常少。

2 词的分布式表示

One-Hot表示仅仅将词符号化，不包含任何语义信息。Harris在1954年提出的“分布假说”为这一设想提供了理论基础：上下文相似的词，其语义也相似。Firth

在1957年对分布假说进行了进一步阐述和明确：词的语义由其上下文决定。

Word Embedding正是这样的模型，而Word2Vec则是其中的一个典型，Word2Vec包含两种模型，即CBOW模型和Skip-gram模型（如下图所示）。以CBOW模型为例，如果有一个句子“the cat sits one the mat”，在训练的时候，将“the cat sits one the”作为输入，预测出最后一个词是“mat”。

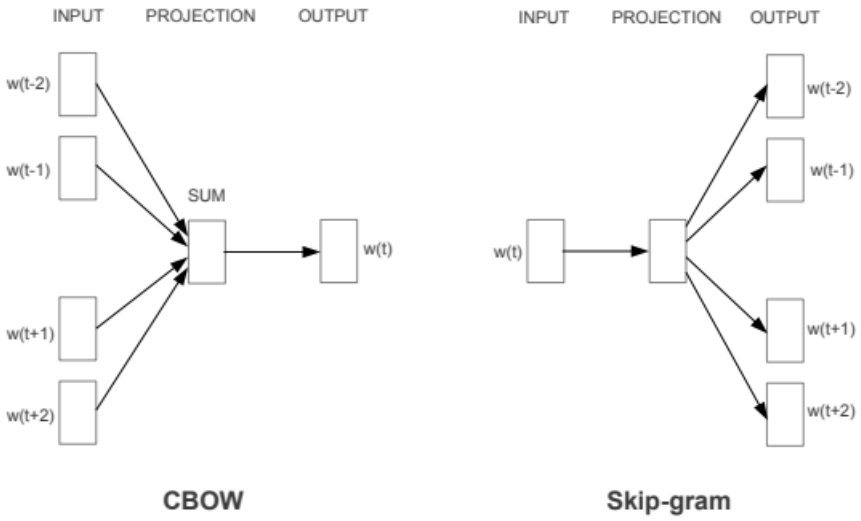

图1 CBOW模型和Skip-gram模型

三、实验

1. 数据预处理

与前两次作业类似，实验前需对实验数据进行预处理。从给定的16本小说语料库中以utf8编码格式读取文件内容后，删除语料库内的所有非中文字符，以及和小说内容无关的片段，得到字符串形式的语料库，然后使用jieba分词进行分词，并使用停用词表进行停用词的过滤，最终返回小说的分词列表。

为了方便后续分析词语之间的相关性，本次实验选取了《射雕英雄传》，《神雕侠侣》，《天龙八部》，《笑傲江湖》，《倚天屠龙记》五本小说，将分词列表保存在txt格式文件中，方便后续使用。

2. 训练Word2Vec模型

使用python库gensim中的Word2Vec类进行模型的训练，并选取5本小说中的代表性人物，分析训练后与该人物相关性最强的10个词：

model = Word2Vec(sentences=PathLineSentences(DATA_PATH), hs=1, min_count=10, window=5, vector_size=200, sg=0, workers=16, epochs=200)

模型的参数释义如下：

sentences：可以是一个list，此处使用的PathLineSentences为一个文件夹下的所有文件，另外有LineSentence对单个文件生效；

hs：如果为1则会采用hierarchica·softmax技巧。如果设置为0（default），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；

window：表示当前词与预测词在一个句子中的最大距离是多少；

vector_size：是指特征向量的维度，默认为100，大的size需要更多的训练数据，但是效果会更好；

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；

workers：线程数；

epoches：训练迭代轮数。

3. K-means聚类

为了进一步验证词向量的有效性，使用TSNE将训练得到的模型中的词向量进行降维（方便展示效果），并使用K-means算法进行聚类。聚类用到的词为5本小说中的代表性人物。最终用散点图进行效果展示。

四、实验结果与分析

1. 相关性分析

实验过程中选取的五本小说以及对应的人物如下:《射雕英雄传》——郭靖，《神雕侠侣》——杨过，《天龙八部》——段誉，《笑傲江湖》——令狐冲，《倚天屠龙记》——张无忌，分析结果如下图所示。

序号	郭靖	杨过	段誉	令狐冲	张无忌
1	黄药师 0.722964	小龙女 0.707415	萧峰 0.626589	岳不群 0.762981	周芷若 0.737059
2	黄蓉 0.716872	黄蓉 0.696542	慕容复 0.618318	林平之 0.680619	赵敏 0.702550
3	欧阳锋 0.700620	李莫愁 0.665691	虚竹 0.605476	岳夫人 0.666960	张翠山 0.686444
4	穆念慈 0.673530	郭靖 0.641884	王语嫣 0.591431	韦小宝 0.660016	谢逊 0.677869
5	洪七公 0.658979	周伯通 0.636734	木婉清 0.581983	岳灵珊 0.659877	金花婆婆 0.630401
6	周伯通 0.656858	洪七公 0.612776	阿朱 0.562739	田伯光 0.658308	俞莲舟 0.613046

7	梅超风 0.649689	陆无双 0.600903	乔峰 0.559085	盈盈 0.632562	蛛儿 0.591263
8	欧阳克 0.643716	欧阳锋 0.594450	段正淳 0.540204	仪琳 0.619162	灭绝师太 0.587697
9	杨过 0.641884	赵志敬 0.592403	鸠摩智 0.531651	任我行 0.617098	韦小宝 0.581226
10	小龙女 0.612617	黄药师 0.592346	钟灵 0.501408	黑白子 0.576637	宋青书 0.567906

根据结果可以看出，词向量相似度较高的词在小说中也有一定关系，以令狐冲为例，岳不群是令狐冲的师傅；林平之是令狐冲的同门师兄弟；岳灵珊是岳不群的女儿，从小和令狐冲一起长大，是令狐冲的师妹；其他人也和令狐冲有一定关系。其他小说中相关分析的结果也较为合理，可以看出词向量的效果较好。

2. 聚类分析

选取5本小说中的几位典型人物，使用K-means聚类的结果如下图所示：

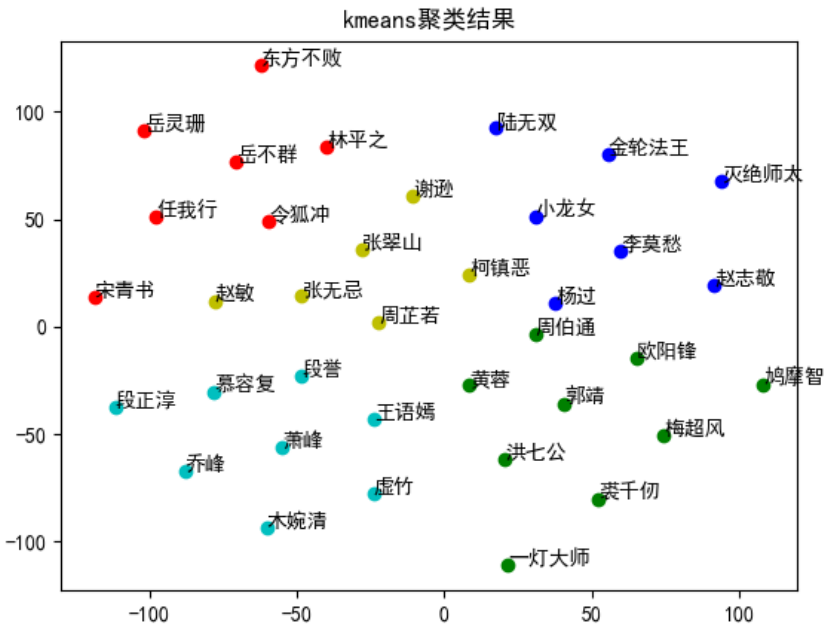

图2 K-means聚类结果

可以看出，同一本小说中的人物基本被分到了同一类中，但也有极少数划分错的情况，比如柯镇恶，但效果总体较好。