Report of Deep Learning for Natural Langauge Processing

Homework_3

Pan Yao
1239388514@qq.com

Abstract

利用给定语料库（金庸语小说料如下链接），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

Word2Vec 是一种流行的词嵌入（Word Embedding）方法，由 Tomas Mikolov 在 Google 开发。它通过将文本中的词映射到向量空间中，使得在这个空间里，语义上相似的词彼此接近，从而能够捕捉到词之间的语义和语法关系。Word2Vec 模型有两种主要的架构：Skip-Gram 和 Continuous Bag of Words (CBOW)。

Skip-Gram 模型的目标是从目标词预测上下文。给定一个特定的目标词，模型预测它周围的上下文词。这种方法尤其有效于处理少见的词汇。

工作流程：

对每个目标词，模型查看其前后一定范围内的上下文词。

模型的输入是目标词的 one-hot 编码，输出是上下文词的概率分布。

使用 softmax 函数将输出层的分数转换成概率。

CBOW 模型与 Skip-Gram 相反，它的目标是根据上下文预测目标词。对于大型数据集，这种方法通常更快并对频繁出现的词汇效果更好。

工作流程：

该模型取一个词的上下文作为输入。

输入层为上下文词的 one-hot 编码，这些编码被平均在隐藏层。

输出层是目标词的预测概率分布，使用 softmax 函数进行概率转换。

训练

Word2Vec 模型的训练通常使用反向传播（Backpropagation）和梯度下降。

本文利用 Word2Vec 模型来训练词向量，通过计算词向量之间的语意相似度、某一类词语的聚类、词语类比来验证词向量的有效性。

Experiment

实验步骤：

1. 准备语料库：本次实验以金庸的 16 部武侠小说作为中文语料库。

2. 预处理：对语料库进行预处理，删除标点符号，无意义的广告等。并使用 jieba 库对文本进行分词。

3. 模型训练：通过 gensim 库中的 Word2Vec 模型对经过预处理的中文语料库进行训练，并通过 model.save 函数保存整个模型。

4. 语意相似度计算。

5. 词类聚类并可视化。

6. 词语类比。

Experimental Results

语意相似度计算：

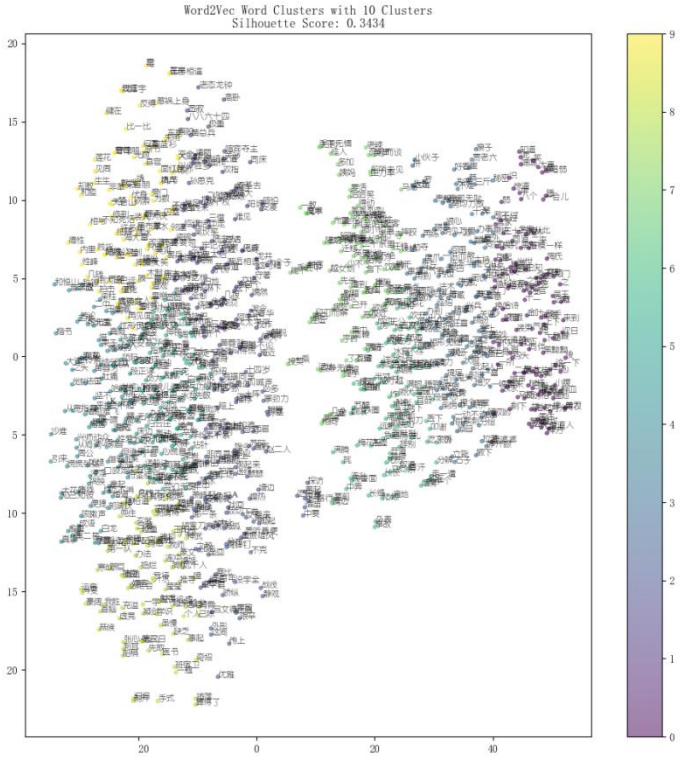
本次实验计算以下对比词语的语意距离：（王，皇帝），（武林，江湖），（冰霜，酒杯）。本次实验在 vector_size=100，epochs=30 的条件下进行语意距离的计算，其结果如下表所示：

对比词	距离
（皇上，皇帝）	0.825255274772644
（武林，江湖）	0.7694724798202515
（冰霜，酒杯）	0.4091034531593323

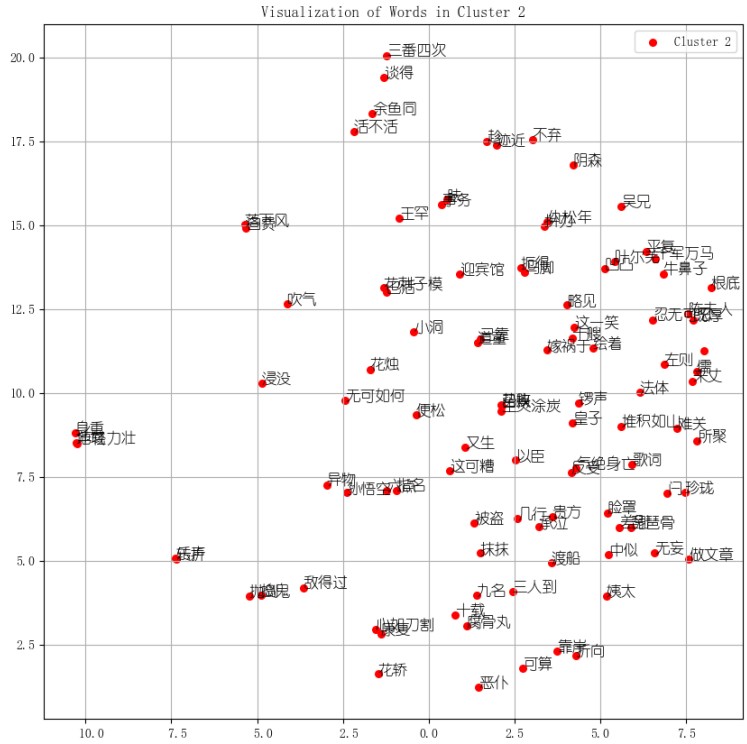
表 1 不同对比词的相似度

词类聚类：

本次实验采用 K-Means，n_clusters=10 进行聚类：



其中不同的颜色代表不同的簇，轮廓系数为 0.3434。



该图为簇索引为 2 的聚类图，其轮廓系数为 0.5326。

该簇的部分词汇为：’无妄’，’马脚’，’中似’，’歌词’，’凹凸’，’余鱼同’，’平复’，’扼得’，’已活’，’十载’，’根底’，’身重’，’心如刀割’，’拚力’，’生灵涂炭’，’锣声’，’抛剑’，’三人到’，’趁’，’贵方’，’孙悟空’，’迎宾馆’，’迹近’，’抹抹’，’这一笑’，’地域’，’事务’，’略见’，’王罕’，’阴森’，’花刺子模’，’落下风’，’道童’，’浸没’，’白费’，’姨太’，’儒’，’小洞’，’六点’，’指名’，’异物’，’承泣’，’被盗’，’气绝身亡’，’肤’，’活不活’，’已靠’，’陈夫人’，’吹气’，’千军万马’，’康复’，’九名’，’腐骨丸’，’叶尔羌’，’既厚’，’琵琶骨’，’左则’，’花烛’，’反受’，’来丈’，’已败’，’仇松年’，’无可如何’，’三番四次’，’皇子’，’渡船’，’忍无可忍’，’靠岸’，’花轿’，’堆积如山’，’嫁祸于’，’谈得’，’敌得过’，’捣鬼’，’差别’，’年轻力壮’，’乐声’，’闩’，’绘着’，’几行’，’吴兄’，’十艘’，’势挟’，，’牛鼻子’，’珍珑’，’脸罩’，’不弃’，’法体’，’可算’，’又生’，’以臣’，’所聚’，’做文章’，’转折’，’恶仆’，’难关’，’这可糟’，’折向’，’便松’

由轮廓系数和部分词汇可以看出该聚类效果并不算很理想，有待改进。

词语类比：

通过两个词列表，其中一个正向加权，一个负向加权，指定两个正向词和一个负向词，从而找到最相似的词。本次实验所采用的词语对：（positive:女人，皇帝，negative:男人），（positive:武林，江湖，negative:侠客），（positive:马蹄，青石板，negative:黑衣）

词	相似词	相似度

（positive:女人，皇帝，negative:男人）	大元帅	0.8061955571174622
（positive:武林，江湖，negative:侠客）	遭遇	0.798683226108551
（positive:马蹄，青石板，negative:黑衣）	锵然	0.8237448930740356