DNLP 作业 3

词向量语义关联探究

张乃天 SY2302516

freshznt@buaa.edu.cn

Abstract

本次作业旨在利用 Word2Vec 模型将文本词语建模成词向量，并利用探究语义关联方法验证词向量的有效性。验证的方法包括语义距离计算、聚类可视化。

In this assignment, we aim to model textual words into word vectors using Word2Vec model. Further, we try to validate the effectiveness of these word vectors by exploring their semantic relationships, including methods of semantic distance computation and clustering visualization.

Introduction

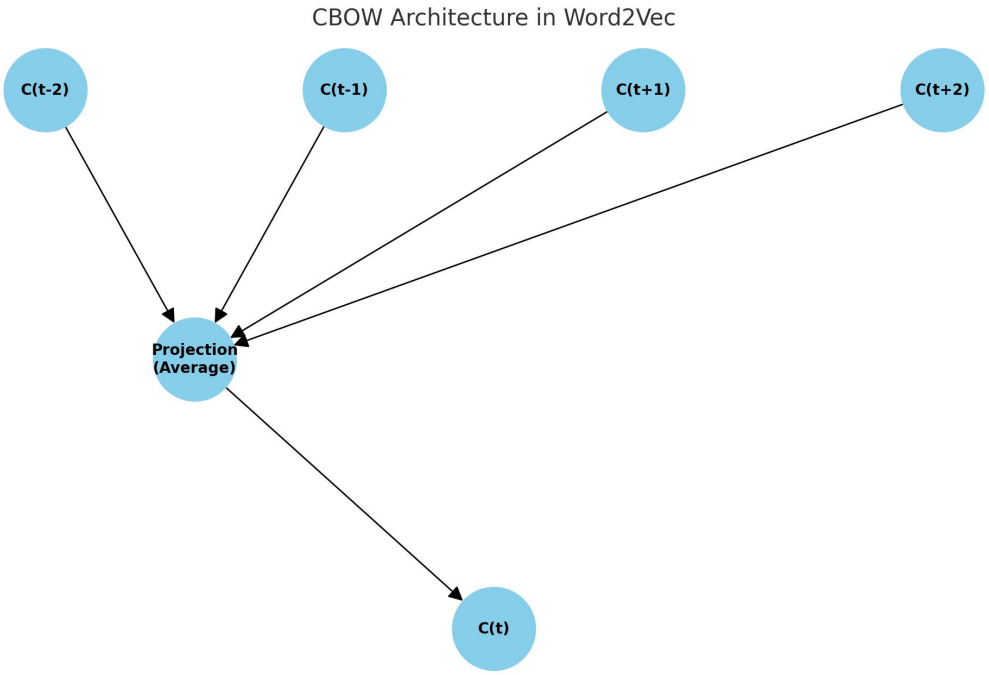
利用给定语料库（金庸语小说料如下链接），利用 1～2 种神经语言模型（如：基于 Word2Vec，LSTM，GloVe 等模型）来训练词向量，通过计算词向量之间的语意距离、某一类词语的聚类、某些段落直接的语意关联、或者其他方法来验证词向量的有效性。

本次作业将采用 Word2Vec 模型，下面简要介绍。

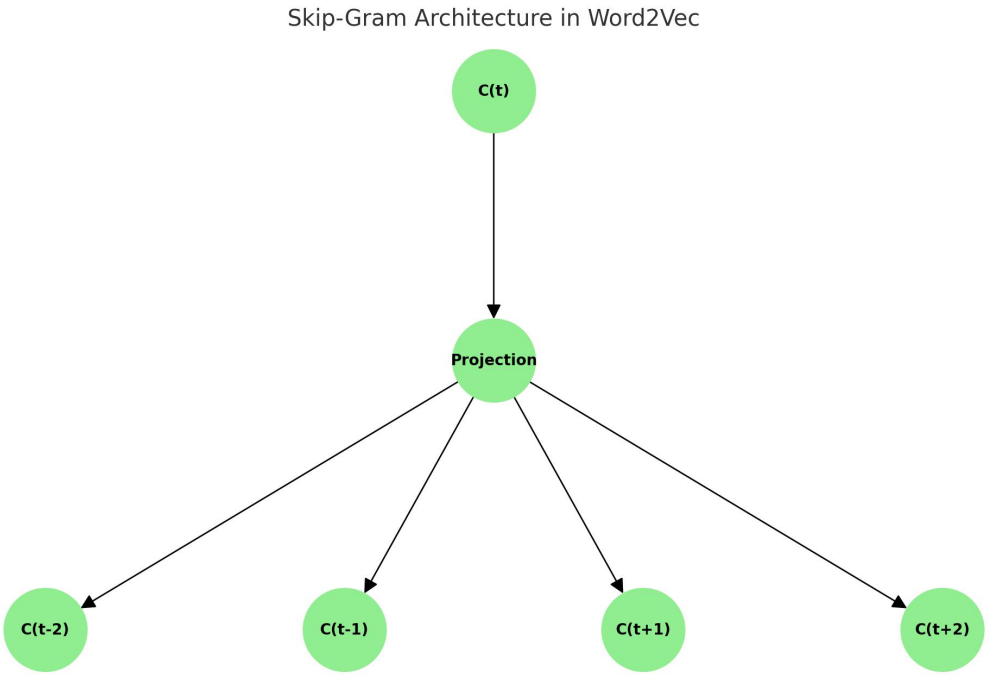
Word2Vec 模型简介：

Word2Vec 是一种广泛使用的从文本数据中学习词向量模型，在各种 NLP 任务中非常有用，例如文本分类、情感分析、机器翻译等。它们提供了一种密集的词表示，捕捉句法和语义相似性，使其成为许多现代 NLP 流程的基础。Word2Vec 主要目标是将单词表示为高维空间中的连续向量，并希望其中具有相似意义的单词彼此接近。目前的 Word2Vec 主要有两种架构：

1）Continuous Bag-of-Words (CBOW)：CBOW 架构基于周围的上下文单词，取上下文单词向量的平均值，并使用它来预测目标单词。这种架构通常更快，适用于较小的数据集。



2）Skip-Gram：Skip-Gram 架构通过给定目标单词预测其周围的上下文单词。具体来说，该架构以单个单词为输入，尝试预测其在定义窗口内的邻近单词。能力：这种架构对较大的数据集特别有效，并且擅长捕捉远距离单词之间的语义关系。



本文将比较研究两种 Word2Vec 模型架构。

Methodology

为了验证词向量的有效性，要在语义层面上分析词向量之间的关系。本次作业采用三种方式来验证词向量之间的语义关联，分别为语义距离计算、聚类可视化以及词语关联性类比

验证。

语义距离计算：词向量语义距离有多种计算方式。例如，余弦相似度衡量的是两个词向量夹角的余弦值，若余弦相似度趋于 1，则两个词夹角很小（语义接近），若余弦相似度趋于 0，则两个词正交（语义较远）；除此之外还有各种范数定义下的距离，例如曼哈顿距离（1 范数）或欧式距离（2 范数），距离越小反应两个词的语义越相近。为了简便，本次作业选择使用余弦相似度衡量语义距离，可以直接使用函数 Word2Vec 类中的 wv.similarity 返回两个词向量的余弦相似度，或者使用 wv.most_similar 返回和目标词汇余弦相似度最相近的几个词语，十分方便。

聚类可视化：可以事先通过人为观察把词义相近的词组成一个组，构建若干个组后使用 Kmeans 聚类方法对这些词对应的词向量进行聚类，并可视化聚类结果。我们希望聚类的结果和人为语义分组结果相近，这样能够显示词向量的有效性。由于整个语料库的词语过于庞大，全部可视化的视觉效果并不好且计算开销过大，因此本次作业人为筛选出三种类别的词语，包括“人名”、“武功”、“地名”，并将 Kmeans 的类别超参数设置为 3。我们期望聚类结果与人为分组的结果相似。可视化时使用 PCA 方法对高维的词向量进行降维，便于在二维图像上可视化聚类结果。

Experimental Studies

Word2vec 模型的超参数设置如下：词向量维度为 20、窗长为 5、负采样个数为 10、最小丢弃次数为 5。

1）语义距离：

以“郭靖”这个词为研究对象，计算与之余弦相似度最接近的 10 个词，其中表 1.1 运用了 CBOW 架构，表 2.2 运用了 Skip-gram 架构。

表 1.1 CBOW 架构下与“郭靖”余弦相似度最接近的 10 个词

排序	词	余弦相似度
1	黄蓉	0.9098
2	杨过	0.8733
3	张无忌	0.8539
4	赵敏	0.8465
5	胡斐	0.8000
6	袁承志	0.7953

7	萧峰	0.7940
8	郭襄	0.7904
9	武三通	0.7662
10	俞岱岩	0.7603

表 1.2 Skip-Gram 架构下与“郭靖”余弦相似度最接近的 10 个词

排序	词	余弦相似度
1	耶律齐	0.9407
2	黄蓉	0.9213
3	武氏	0.9132
4	武三通	0.9104
5	杨过	0.9049
6	朱二人	0.8845
7	金轮法王	0.8719
8	黄药师	0.8679
9	张无忌	0.8603
10	殷梨亭	0.8596

可以看出，余弦相似度前十的词语都是小说角色名字，验证了词向量的有效性。

2）聚类可视化：

为了给出聚类的可视化结果，事先从语料库中选出三类词语：人名、武功和地名（人名：'郭靖','黄蓉','杨过','赵敏','张无忌','张翠山','郭襄','小龙女','武三通','周芷若'；武功：'一阳指','六脉','降龙十八掌','打狗棒法','乾坤','太极拳','神剑','蛤蟆功'；地名：'衢州','天津','汴梁','信阳','广州','广西','凉州','白马寺','镇江','太原'）。下面给出高维词向量在聚类之后，经过 PCA 降维之后的二维图像结果。其中图 1.1 是 CBOW 架构下的聚类结果，图 1.2 是 Skip-Gram 架构下的聚类结果。

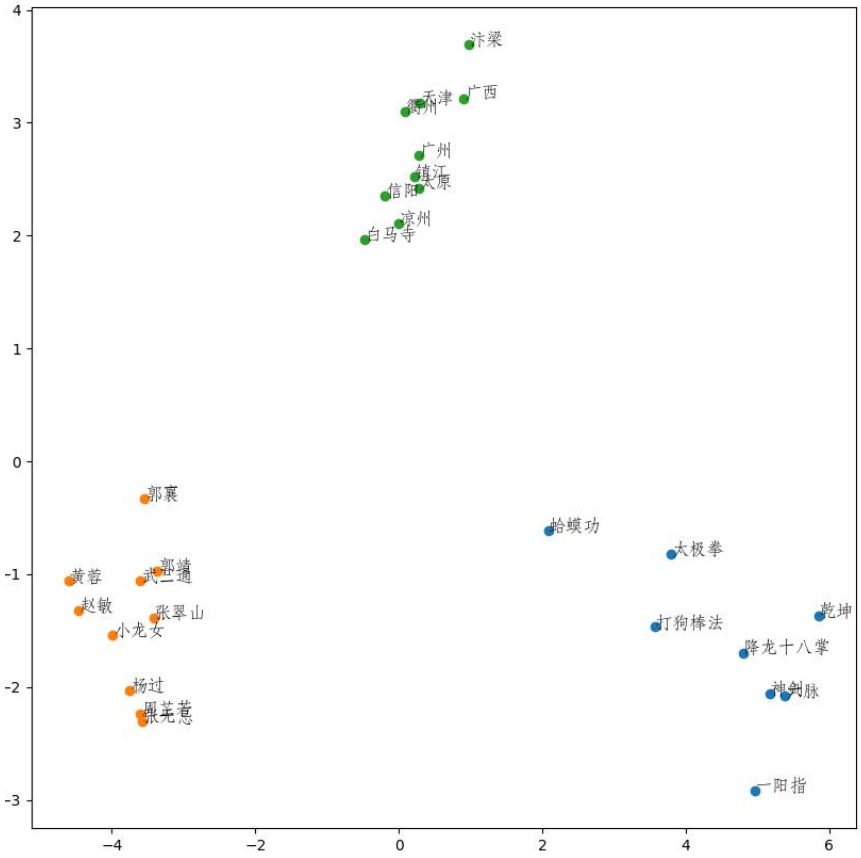


图 1.1 CBOW 架构下词语聚类可视化结果

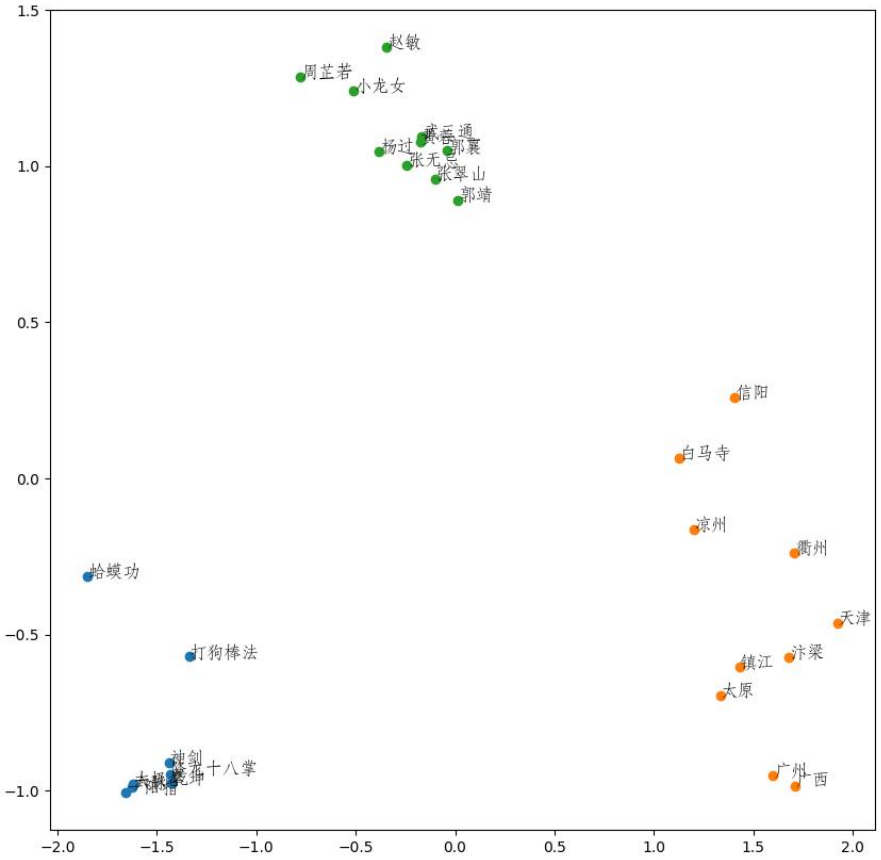


图 1.2 Skip-Gram 架构下词语聚类可视化结果

由图 1.1 和 1.2 可以看出，无论是 CBOW 还是 Skip-Gram 架构，聚类结果把三种不同类别的词语准确的分成了三个对应组别，验证了词向量的有效性。

Conclusions

本次作业主要探究了 Word2Vec 模型建模的词向量的有效性。通过语义距离和聚类可视化的结果可以看出，Word2Vec 模型的两种架构（CBOW 和 Skip-Gram）所建模的词向量有效性较好，比较符合语料库中的语义。