Report of Deep Learning for Natural Langauge Processing

Homework 3

Ao Xie
xieao2019@buaa.edu.cn

Abstract

本研究通过对 16 部小说的中文语料库进行预处理作为数据集，使用 Word2Vec 模型训练词向量，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，来验证词向量的有效性

Introduction

词向量（Word Embeddings）是一种将词语映射到高维向量空间的方法，使得语义相似的词语在向量空间中彼此接近。这种方法在自然语言处理（NLP）领域有着广泛的应用，包括文本分类、情感分析、机器翻译等。

主要有两种常见的词向量模型：

静态词向量模型：如 Word2Vec、GloVe。这些模型通过大规模语料库的训练，生成固定的词向量。这意味着一个词在不同的上下文中具有相同的向量表示。

动态词向量模型：如 ELMo、BERT。这些模型生成的词向量是上下文相关的，即同一个词在不同的句子中可能有不同的向量表示。

常见的词向量模型：

1. Word2Vec

Word2Vec 是由 Google 提出的词向量模型，主要有两种训练方法：

Skip-Gram：目标是预测给定中心词的上下文词。即通过中心词预测其附近的词。

CBOW（Continuous Bag of Words）：目标是通过上下文词预测中心词。例如，给定句

子 "I love natural language processing"：在 Skip-Gram 模型中，"love" 可以作为中心词，目标是预测 "I"、"natural"、"language"、"processing"；在 CBOW 模型中，"I"、"natural"、"language"、"processing" 可以作为上下文词，目标是预测中心词 "love"。

2. GloVe

GloVe（Global Vectors for Word Representation）是由斯坦福大学提出的模型。它通过构建全局词共现矩阵并进行矩阵分解来生成词向量。GloVe 的目标是通过捕捉全局统计信息来更好地表示词语的语义。

3. FastText

FastText 是由 Facebook 提出的词向量模型，它的主要改进在于将词语拆分成字符 n-gram，从而可以更好地处理未登录词（OOV）。例如，词语 "running" 可以被拆分为 "run"、"runn"、"running" 等字符 n-gram。

4. BERT

BERT（Bidirectional Encoder Representations from Transformers）是由 Google 提出的预训练语言模型，它通过双向 Transformer 编码器来捕捉词语在上下文中的表示。BERT 的一个显著特点是它能够生成上下文相关的词向量。

应用场景方面，词向量可以应用在：

1. 文本分类

词向量可以用作文本分类任务中的输入特征。通过将句子中的每个词转换为词向量，然后对这些向量进行平均或其他聚合操作，可以生成句子的向量表示，从而用于分类任务。

2. 情感分析

在情感分析中，词向量可以帮助捕捉文本中的情感信息。例如，词语 "happy" 和 "joy" 在词向量空间中会非常接近，而 "sad" 和 "happy" 则会相距较远。

3. 机器翻译

词向量在机器翻译中也有广泛应用。通过将源语言和目标语言的词语转换为向量表示，机器翻译模型可以更好地捕捉不同语言之间的对应关系。

4. 语义搜索

在语义搜索中，词向量可以帮助提高搜索结果的相关性。通过将查询和文档转换为词向量表示，可以计算它们之间的相似度，从而返回更相关的搜索结果。

词向量是自然语言处理中的一种重要技术，通过将词语表示为高维向量，能够有效捕捉词语之间的语义关系。常见的词向量模型包括 Word2Vec、GloVe、FastText 和 BERT，它们在各种 NLP 任务中有着广泛的应用。通过不断优化和改进词向量模型，可以提高 NLP 应用的准确性和效果。

Methodology

具体方法方面，共分为数据集预处理、训练模型、分析语义相似度、聚类分析、降维和绘制图像。

1.数据集预处理，将 16 部小说去除停顿词，分词并合并成一个文本，作为数据集。

2.训练模型，使用 gensim 库的 Word2Vec 模型对处理后的语料库进行训练。

3.词语相似度计算：计算两个词语之间的语义相似度。

4.KMeans 聚类分析：对词向量进行 KMeans 聚类。

5.TSNE 降维和绘制图像：对一簇词向量使用 PCA 初始降维和 TSNE 降维，并绘制散点图。

Experimental Studies

1 数据集预处理

实验使用的数据集包括 16 部中文小说，首先去除停顿词，然后使用 Jieba 分词，合并成一个文本，作为数据集。

2 实验结果

实验中，首先利用 train_word2vec_model 进行训练，向量维度为 100，窗口长度设置为 5，共训练 100 轮。过程中使用多核并行处理，降低训练时间。

挑选有代表性的几对词语，进行 sklearn 库中的函数进行语义相似度计算，结果如下：

表 1 几对词语的语义相似度

词对	父亲/爹爹	儿子/女儿	师父/师尊	杨过/小龙女	倚天剑/屠龙刀	武当/少林
相似度	0.972	0.989	0.080	0.029	0.005	0.004

对词向量进行聚类操作，并随机抽取部分样本（500 个）进行可视化，将其中一簇使用 PCA 和 TSNE 进行降维，绘制散点图如下。

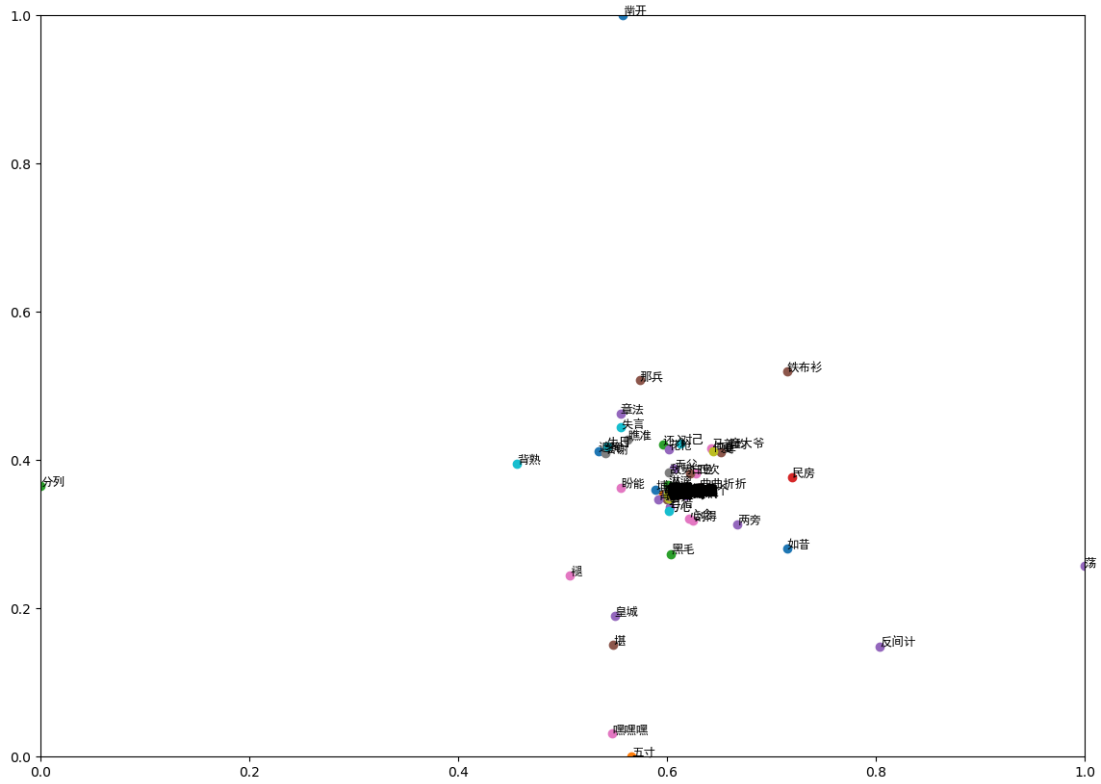

图 1 其中一簇的散点图

Conclusions

实验结果显示，语义相近的词语通过词向量计算的语义相似度较高，其中“父亲”和“爹爹”、“儿子”和“女儿”的相似度接近 1；而“师父”和“师尊”指代的不一定是同一人，使用语境和场合也不尽相同，因此相似度中等；“杨过”和“小龙女”虽为不同人物，但由于在小说中为情侣，相似度也中等；而“倚天剑”与“屠龙刀”、“武当”与“少林”为同一类别不同物品或门派，语境语义差距比较大，语义相似度低。聚类图显示语义相近的词语被归为一簇，其中“遮掩”、“失言”等词语明显具有同类词特征，与预期相符。综上所述，本研究验证了词向量的有效性。