Report of Deep Learning for Natural Langauge Processing 3

Yongquan Chen
cyqss@buaa.edu.cn

Abstract

在这个任务中，我们探索了 Word2Vec 在使用金庸小说语料库训练词嵌入中的应用。语料库经过预处理，以在标记化之前删除标点符号和停用词。然后，本文训练 Word2Vec 模型来生成词向量，它捕获了单词之间的语义关系。为了验证经过训练的词向量的有效性，我们计算了所选单词之间的语义距离，对相似单词进行聚类，并分析了段落之间的语义关联。此外，使用了 t-SNE 可视化词向量以降低维数，从而提供词嵌入的清晰图形表示。这种方法展示了 Word2Vec 在自然语言文本中发现和表示语义结构的能力，使其成为各种自然语言处理任务的强大工具。

Introduction

Word2Vec 是一种用于生成词向量的算法，由 Google 的 Tomas Mikolov 等人提出。其基本思想是将词语映射到高维向量空间中，使得在语义上相似的词语在向量空间中也彼此接近。Word2Vec 主要有两种模型架构：CBOW 和 Skip-Gram。CBOW 模型通过上下文词预测目标词，而 Skip-Gram 模型则通过目标词预测其上下文词。在实际应用中，Word2Vec 通过神经网络训练，在大规模文本语料上学习到每个词的低维向量表示。这些向量可以捕捉到词语之间的语义关系，例如“国王”减去“男人”再加上“女人”接近“王后”。这种能力使 Word2Vec 在自然语言处理任务中得到广泛应用，如文本分类、情感分析和信息检索等。

t-SNE 是一种用于数据降维和可视化的技术，由 Laurens van der Maaten 和 Geoffrey Hinton 提出。t-SNE 特别适用于高维数据的可视化，它能将高维数据嵌入到低维（通常是 2D 或 3D）空间中，同时保留数据的局部结构。t-SNE 通过将高维数据点间的相似度转换为低维空间中的概率分布，并最小化两者之间的 Kullback-Leibler 散度，使得在高维空间中相似的数据点在低维空间中也彼此接近。这使得 t-SNE 在可视化复杂数据集时，能够揭示出数据的群体结构和模式。

Methodology

对给定的金庸数据集，本文首先将这些文本合并为一个大文本文件。接着，使用 Jieba 库对文本进行分词，并去除标点符号和停用词，以获得干净的词语列表。然后，利用 Gensim 库中的 Word2Vec 模型对预处理后的文本数据进行词向量训练，通过选择合适的参数来优化模型效果。

在训练完词向量后，本文计算了若干关键词之间的余弦相似度，以验证词向量是否能够正确反映词语间的语义关系。接着，通过 KMeans 等聚类算法，对词向量进行聚类，观察同

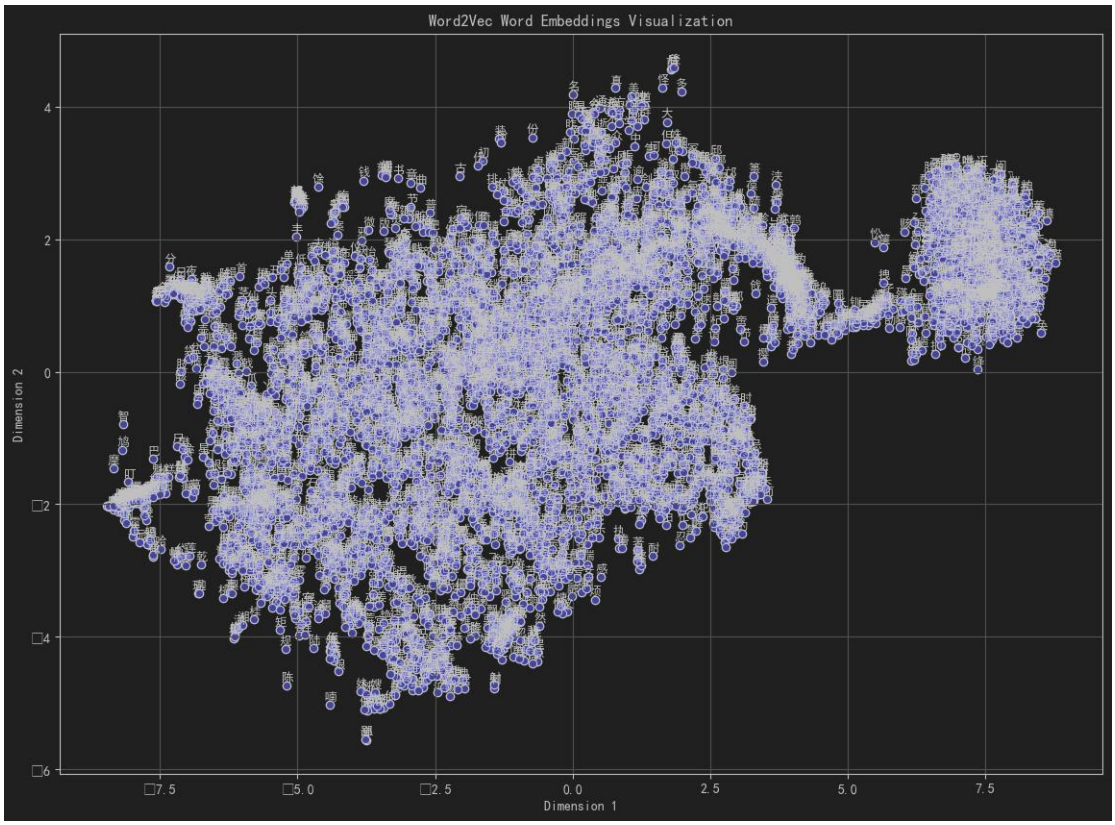
一语义类词语是否能够被正确地聚在一起。此外，本文还通过将每个段落表示为其包含词语的词向量平均值，计算了段落向量之间的余弦相似度，分析段落之间的语义关联。

最后，使用 PCA 和 t-SNE 对高维的词向量进行降维，并使用 Matplotlib 库对降维后的词向量进行可视化，直观展示词语之间的语义关系和聚类情况。通过这些步骤，本文展示了 Word2Vec 在中文文本语义分析中的有效性，并通过可视化手段增强了对词向量语义结构的理解。 模型。最后用训练好的模型计算两个词向量的相似度。进行交叉验证。最后输出分类准确率。

Experimental Studies

M1: t-SNE 进行降维可视化

对 Word2Vec 模型进行降维可视化:



M2:段落的词向量比较

对四组段落进行比较:

金庸选段："这一阵歌声传入湖边一个道姑耳中。她在一排柳树下悄立已久，晚风拂动她杏黄色道袍的下摆，拂动她颈中所插拂尘的千百缕柔丝，心头思潮起伏，当真亦是'芳心只共丝争乱'。"

古龙选段："天涯远不远？不远！人就在天涯，天涯怎么会远？明月是什么颜色的？”“是蓝的，就像海一样蓝，一样深，一样忧郁。”“明月在哪里？”“就在他心里，他的心就是明月。”“刀呢？”“刀就在他手里！”“那是柄什么样的刀？”“他的刀如天涯般辽阔寂寞，如明月般皎洁忧郁，有时一刀挥出，又仿佛是空的！"

刘慈欣选段："章北海慢了一步，尽管他执剑至生命最后一刻，但依然死于同类之手。可是章北海不后悔，相反他感到解脱，因为自己的终极使命已经完成，人类的火种最终继续

延续。章北海说出了令所有人为之动容的一句话：没关系的，都一样。"

唐三选段："超过五万年修为的魂兽数量虽然也不多，但却绝不像十万年魂兽那么稀少。不过，经历了唐三这次洗劫，也几乎将落日森林内五万年以上修为的魂兽消灭了大半。除了蓝银皇恢复到九环之外，他的昊天锤也吸收了两个魂环。所有魂环，全部是五万年以上修为。"

段落相似度矩阵：

	0	1	2	3
0	1.000000	0.277506	-0.047067	0.019844
1	0.277506	1.000000	0.088804	0.245144
2	-0.047067	0.088804	1.000000	0.158002
3	0.019844	0.245144	0.158002	1.000000

观察到古龙和金庸的相似度更高，其他人相似度没有这么高。