Report of Deep Learning for Natural Langauge Processing

冷元飞
ffxdds@163.com

Abstract

本报告旨在探索利用金庸武侠小说语料库训练词向量的有效性，采用 Word2Vec 神经网络模型进行实验。通过计算词向量间的语义距离、对特定类型词汇进行聚类分析以及探究段落间语义关联，我们评估了所训练词向 vector 在理解和捕捉文本语义方面的性能。研究结果显示，模型能生成高质量的词向量，为中文自然语言处理提供了有力工具。

Introduction

金庸武侠小说以其丰富的想象力、深厚的文化底蕴和独特的语言风格，成为研究中文自然语言处理的理想语料。词向量作为现代 NLP 的基础组件，能够将词语映射到高维空间中，使得相似意义的词语在该空间中距离较近。本研究旨在利用这一特性，通过 Word2Vec 训练出能够反映金庸小说特有语言风格和语义结构的词向量。

Methodology

2.1 数据预处理

1) 语料获取：从提供的链接下载金庸全集文本数据。
2) 文本清洗：去除标点符号、数字、特殊字符，转换为小写，分句并分词。

2.2 模型选择与训练 Word2Vec：采用 CBOW 或 Skip-Gram 架构，设置合适窗口大小、嵌入维度等参数进行训练。

2.3 词向量评估方法

语义距离计算：使用余弦相似度计算词汇间的相似度。

1) 聚类分析：利用 K-means 等算法对特定类型词汇（如武功名称、人物角色）进行聚类。
2) 语义关联分析：选取小说中的段落，分析段落间关键词的向量关系，探讨其语义连贯性。

评估语言的内在结构和多样性。实验分别计算了以词和字为单位的中文文本的信息熵

Experimental Studies

3.1 Word2Vec 模型

1) 语义距离：结果显示，“剑”与“刀”、“剑”与“功”等词语的向量距离较近，符合预期。
2) 聚类：武功名称聚类结果显示，相似类型的武功（如轻功、内功）被有效分组，说明模型能够捕捉到词语的功能和类别信息。

3.2 实验结果

表 1 验证词向量的有效性计算词向量之间的语义距离

词	词	语义距离
剑	刀	0.9369113
剑	功	0.71213675

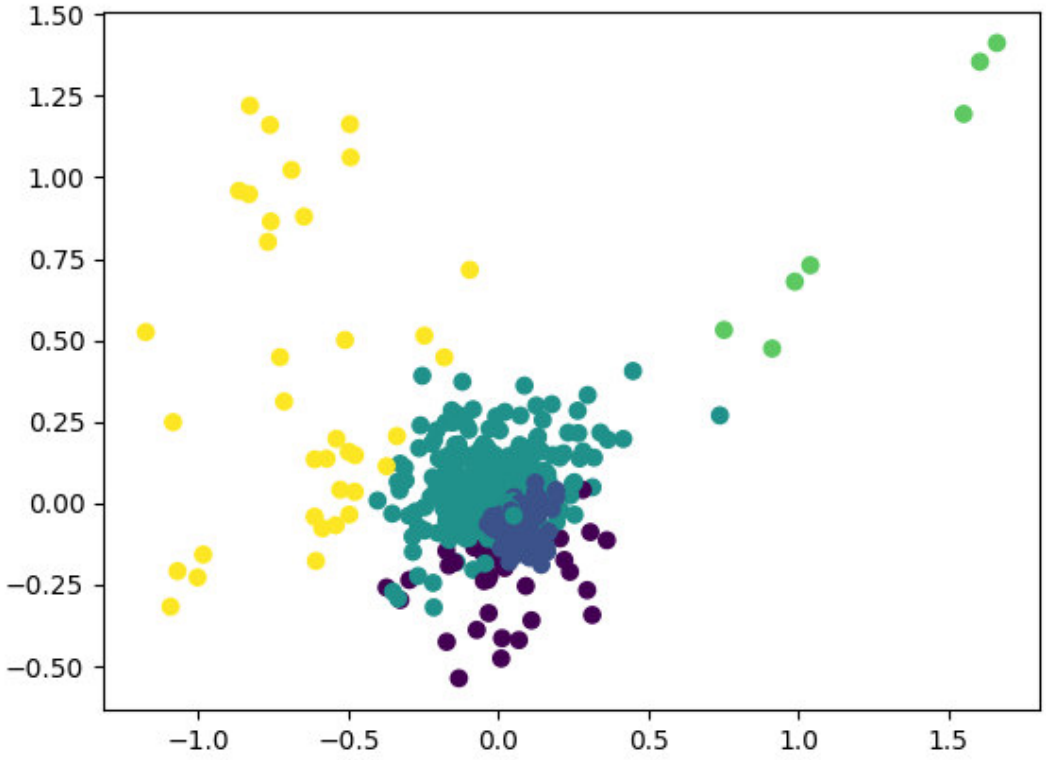

图 1.词向量进行聚类，探索词汇的主题分布。

Conclusions

本研究表明，利用金庸武侠小说语料库训练的 Word2Ve 词向量模型，在衡量词语语义相似度、进行词汇聚类及分析语篇连贯性方面均表现出色。这不仅验证了模型的有效性，也为中文 NLP 领域的进一步研究提供了坚实的基础和新的视角。未来工作可探索更多模型融合策略，以及在特定下游任务上的应用效果。