基于 Word2Vec 的语义分析与词向量可视化

宋雯婧
13501982761@163.com

摘要

本文旨在通过自然语言处理（NLP）技术分析金庸小说《天龙八部》中的语义结构。利用 Word2Vec 神经网络模型训练词向量，并通过 t-SNE 降维方法对词向量进行可视化，从而揭示该小说中的词语间的语义关系。实验结果表明，训练出的词向量能够有效地捕捉到小说中关键人物、地点、武功和门派等概念的语义关联，且通过可视化展示出词语间的聚类情况。通过对关键词语的向量范数检查，验证了词向量的有效性，并通过图形化呈现进一步展示了语义的聚类现象。

1. 引言

金庸的武侠小说以其庞大的世界观、复杂的人物关系和丰富的文化背景吸引了众多读者。《天龙八部》作为其中一部经典作品，涉及大量的人物、地点、武功以及门派，其文本内容蕴含了丰富的语义信息。通过自然语言处理中的词向量模型，我们可以深入挖掘文本的潜在语义结构。

本研究通过训练 Word2Vec 模型对《天龙八部》文本进行分析，探索小说中各个重要元素之间的语义关系，并利用 t-SNE 降维技术对词向量进行可视化展示，直观地展示词语间的聚类效果。

2. 实验方法

2.1 数据与预处理

首先，我们对《天龙八部》的原始文本进行了预处理，主要包括文本清洗和分词。为了有效地处理小说中的专有名词和关键词，我们通过 jieba 分词工具并结合自定义词典对关键词（如人物、武功等）进行了特殊处理，确保这些词语在分词时不被拆分。处理后的文本被分为多个段落，每个段落内的词语被提取并用于训练 Word2Vec 模型。

预处理流程的具体步骤如下：

（1）加载用户字典：通过 jieba.load_userdict()加载自定义词典，以保证特定的专有名词和词汇（如人物、地点等）能够正确分词。

（2）保护关键词：为确保某些关键字（如人物、门派等）在分词过程中不被拆分，使用_protect_keywords()方法为它们添加特殊标记※，处理后再移除这些标记。

2.2 词向量训练

在数据预处理完成后，我们使用 Word2Vec 模型对文本进行训练。Word2Vec 模型通过以下两个主要训练方法来学习词语的向量表示：

（1）连续词袋模型（CBOW, Continuous Bag of Words）：该模型通过上下文词语来预测目标词的词向量。
（2）跳字模型[1]（Skip-Gram）：该模型则相反，通过给定的词语来预测其周围的上下文词语。
在我们的任务中，使用了跳字模型（Skip-Gram），其数学表达式为：

$$p(w_t|w_{t-k},\dots,w_{t+k}) = \prod_{-k\le j\le k, j\neq 0} p(w_{t+j}|w_t)$$

其中，w_t是目标词，w_{t+j}是上下文词，k是窗口大小。

模型的主要参数包括：
Ⅰ词向量维度（vector_size=200）：表示词向量的大小。
Ⅱ窗口大小（window=10）：表示模型在训练时考虑上下文的范围。
Ⅲ最小词频（min_count=2）：过滤掉在文本中出现次数较少的词语。
Ⅳ向量归一化：通过归一化处理向量，确保词向量的稳定性。

2.3 t-SNE 降维与可视化

训练完成后，为了便于理解和展示词向量之间的关系，我们使用 t-SNE 算法将高维词向量降至二维，并利用 Matplotlib 进行可视化。t-SNE 算法通过优化以下目标函数来降低高维数据到低维的映射误差：

$$KL - Divergence = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

其中，p_{ij}是高维空间中数据点i和j之间的相似度，q_{ij}是低维空间中的相似度。

在可视化过程中，我们选取了包含多个关键词的词向量，如“段誉”、“乔峰”、“虚竹”等，并通过 t-SNE 降维[2]后展示其在二维空间中的分布。为了避免标签重叠，我们使用了 adjustText 库进行动态调整，使得每个词语的标签能够清晰可见。

3.实验结果与分析

3.1 词向量的有效性验证

我们对几个关键字（如“段誉”、“乔峰”等）的词向量进行了检查，发现这些关键词已经成功地被模型收录，并且其向量范数接近 1，说明词向量的训练效果较好。每个词语的向量范数均在 1 附近，表明这些词语在语义空间中的表示是有效且合理的。

3.2 可视化结果

通过 t-SNE 降维，我们将训练得到的词向量映射到二维空间，图中的不同点代表了不同的词语，点之间的距离反映了它们在语义上的相似度。我们可以看到，图中的词语明显分成了几个簇，其中“段誉”、“乔峰”和“虚竹”等

主要人物形成一个聚类，而“降龙十八掌”、“天山六阳掌”等武功则组成另一个聚类，此外，“少林寺”和“丐帮”这类门派词语也分布在一起。通过这种可视化，我们可以直观地看到不同词语之间的语义关系。

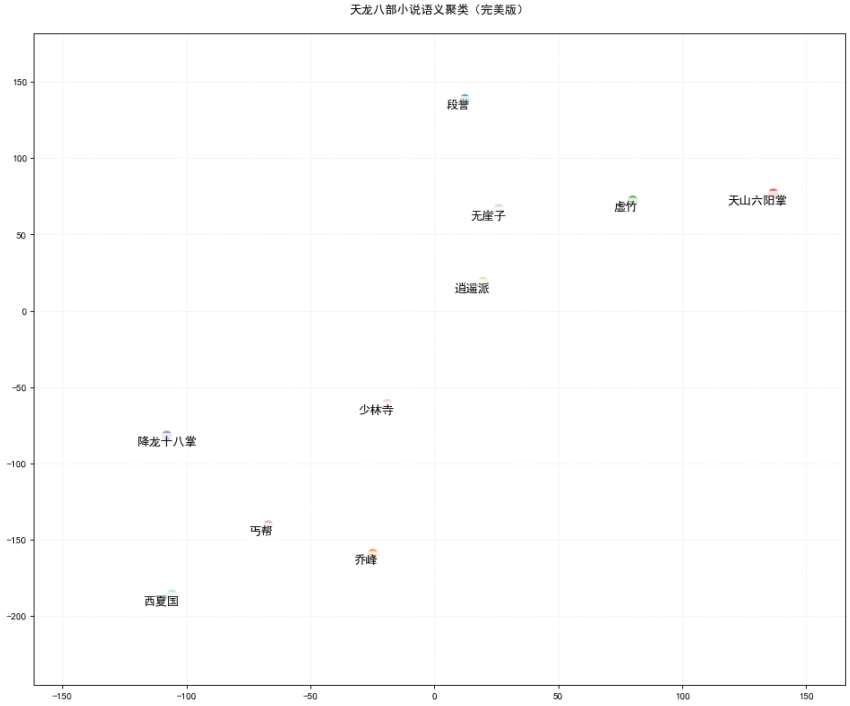

图 1　运行结果图

4.结论

实验结果表明，Word2Vec 模型在捕捉《天龙八部》中的语义关系方面表现出色。通过 t-SNE 降维后的可视化图，我们能够直观地观察到不同类别的词语在语义空间中的分布情况，如人物、武功、门派等。尤其是在关键词汇聚的区域，我们看到了明显的语义聚类现象，这证明了词向量在语义分析中的有效性。

然而，t-SNE 算法在降维过程中可能会受到局部最优解的影响，导致不同类别之间的距离不完全准确。因此，进一步的优化和其他降维方法（如 UMAP）可能会进一步提升结果的准确性。

本文通过 Word2Vec 模型对《天龙八部》进行词向量训练，并使用 t-SNE 进行可视化展示。实验结果验证了词向量模型能够有效地捕捉到文本中的语义关系。通过这种方法，我们能够进一步挖掘文本中的潜在结构，为后续的文本分析和理解提供有力支持。

5.参考文献

[1] Mikolov, T., et al. (2013). Distributed Representations of Words and Phrases and Their Compositionality. *Neural Information Processing Systems*.
[2] Van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9, 2579-2605.