Report of Deep Learning for Natural Langauge Processing

张永易
2394822700@qq.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，具体采用了 Word2Vec 神经网络模型进行一系列实验。金庸的武侠小说以其丰富的词汇和深厚的文化背景而著称，为研究中文自然语言处理提供了独特且富有挑战性的语料库。在实验中，我们通过计算词向量间的语义距离来评估模型的性能，旨在检测模型在多大程度上能够捕捉词汇间的语义相似性。此外，我们还对特定类型的词汇进行聚类分析，以进一步探讨模型在识别和归类相似词汇方面的能力。例如，通过分析“武功”、“人物”以及“地点”等类别的词汇聚类结果，我们能够直观地观察到模型在理解这些类别内词汇的语义关系上的表现。此外，本研究还探讨了段落间的语义关联，通过比较不同段落的词向量表示，评估模型在捕捉段落级别语义方面的能力。这个过程有助于验证模型在处理较长文本时的有效性，特别是在理解和分析复杂故事情节时的表现。研究结果显示，基于金庸武侠小说语料库训练的 Word2Vec 模型能够生成高质量的词向量。这些词向量不仅在词汇语义相似性上表现出色，而且在词汇聚类和段落语义关联分析中也展示了良好的性能。这表明，利用金庸武侠小说这样的丰富中文文本进行词向量训练，可以为中文自然语言处理提供有力的工具，具有广泛的应用前景。通过这项研究，我们不仅验证了 Word2Vec 模型在中文文本处理中的适用性，还为进一步研究和应用打下了坚实的基础。未来的研究可以考虑引入更多类型的中文文本，或采用更为先进的神经网络模型，进一步提升词向量的质量和应用效果。

Introduction

词嵌入模型是一种在自然语言处理中广泛应用的技术，其核心目标是将词汇表示为连续向量，以便更好地捕捉词汇之间的语义关系。在这些模型中，Word2Vec 凭借其简单而高效的特性成为了研究和实践中的焦点。Word2Vec 通过训练神经网络，学习每个词汇的稠密向量表示，使得具有相似语义的词汇在向量空间中彼此靠近，从而为词汇之间的语义关联提供了强大的数学表达。

Word2Vec 模型的有效性在于其在多个任务上的出色表现。首先，它能够准确捕捉词汇之间的语义相似性，使得语义相关的词汇在向量空间中距离较近。其次，通过简单的向量运算，Word2Vec 可以执行词汇类比推理，例如“男人：女人::国王：皇后”，这种能力体现了模型对于词汇语义关系的深刻理解。此外，Word2Vec 模型还能够处理大规模语料库中的稀疏数据，提升了模型的泛化能力和应用范围，从而为各种自然语言处理任务提供了强有力的支持。

Methodology

Word2vec 模型

Word2Vec 是一种经典的词嵌入（word embedding）技术，它由两个主要模型组成：CBoW（Continuous Bag of Words）和 Skip-gram。这两种模型都是用来学习词向量的，即将单词表示成稠密的向量，使得具有相似含义的单词在向量空间中距离较近。

CBoW 模型

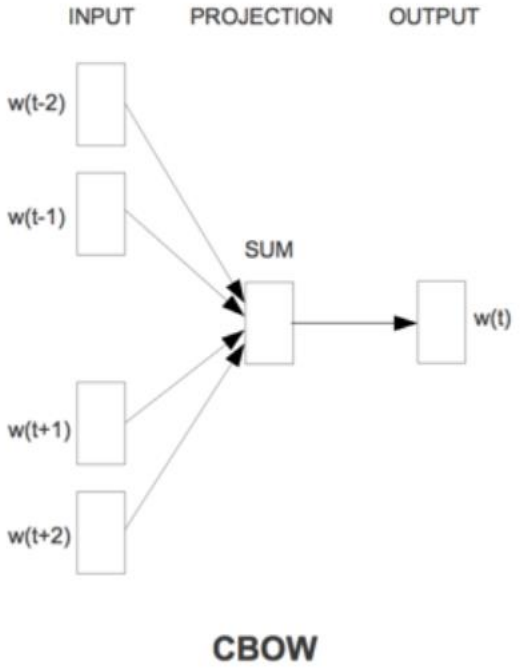

图 1 CBoW 模型结构示意图

CBoW 模型（Continuous Bag of Words）：在 CBoW 模型中，给定一个中心词的上下文窗口（即周围的词），模型的任务是预测这个中心词。具体来说，它通过上下文中的词向量的平均值来预测中心词。

Skip-gram 模型

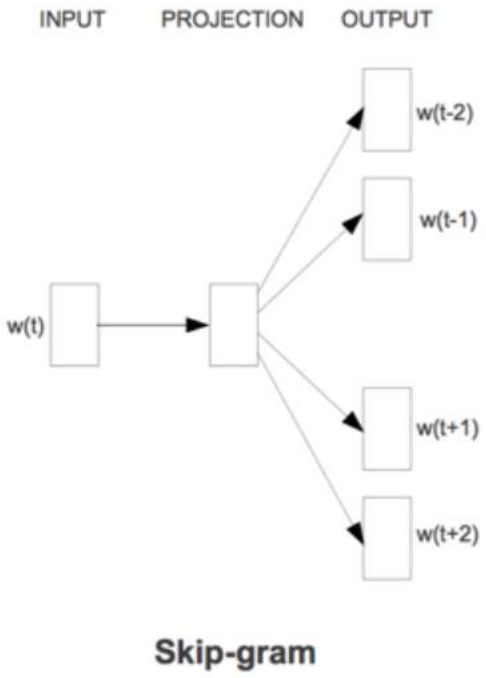

图 2 Skip-gram 模型结构示意图

Skip-gram 模型：与 CBoW 相反，Skip-gram 模型是给定一个中心词，预测它的上下文词。它的任务是在给定中心词的情况下预测周围词的概率分布。

这两种模型都使用了神经网络，通过训练来调整词向量的参数，使得在语料库中经常出现在相似上下文中的词在向量空间中更接近。这样的词向量可以用于许多自然语言处理任务，如语义相似度计算、命名实体识别等。

Experimental Studies

词嵌入模型训练

递归搜索文件夹中的所有文件路径、读取每个小说文件，并将其内容存储在一个语料库列表中、清洗文本数据，去除非中文字符、英文字符、数字和符号，并统计噪声符号的出现次数、根据噪声符号的出现次数，筛选出出现次数低于阈值的噪声符号、遍历语料库中的每一行文本，将噪声符号替换为空字符串、将处理后的文本写入到一个新文件中。

打开预处理后的文本文件，然后逐行读取文本内容。对于每一行文本，使用 jieba 库的 lcut 函数进行中文分词，得到每个词的列表。接着，通过正则表达式过滤掉所有的标点符号，并将分词结果存储在 lines 列表中。

使用了 Gensim 库中的 Word2Vec 模型对分词后的文本进行训练。它通过提供分词后的文本数据来训练 Word2Vec 模型，该模型将文本中的单词转换为向量表示。

词向量有效性分析

从词向量的距离角度可以验证词嵌入模型生成的词向量是否符合预期。如“剑”和“棒”的词向量如下所示：

剑的词向量：

[0.2694778 0.6641987 -0.14718382 1.0386069 -0.4188205 -0.44348332
  0.46155977 -0.2707483 -1.1473352 -0.18541867 0.6744134 -0.58602595
  0.79218554 -0.48213947 -1.1408798 0.86038893 0.46143103 0.19649957
 -0.8823881 -1.4048852]

棒的词向量：

[0.4289369 0.47459677 -0.1723223 0.7232688 -0.34161848 -0.32719707
  0.08355637 0.08376456 -0.8919226 -0.17853133 0.5347331 -0.6642583
  0.39224082 -0.53308344 -0.84366447 0.6556872 0.7568995 -0.1939237
 -0.80557334 -1.4213483]

计算可得“剑”和“棒”的相似度为 0.950791，进一步可以展示与“剑”的词向量最接近的 20 个词向量对应的词，其分别为：

[('棒', 0.9507908821105957),
 ('刀', 0.9199104905128479),
 ('倚天剑', 0.9027681946754456),
 ('金蛇剑', 0.9010294675827026),
 ('拂尘', 0.9005709290504456),
 ('木剑', 0.8997235894203186),
 ('短剑', 0.8917266726493835),
 ('招', 0.887366771697998),
 ('鬼头', 0.8806307315826416),
 ('七伤', 0.8757781982421875),
 ('剑鞘', 0.8737879991531372),
 ('短刀', 0.8725529909133911),
 ('铁剑', 0.8716607689857483),
 ('空手', 0.8699184060096741),
 ('单刀', 0.8694975972175598),
 ('长剑', 0.862953782081604),
 ('飞刀', 0.8617702722549438),
 ('刃', 0.8609581589698792),
 ('重剑', 0.8595697283744812),
 ('拳', 0.8585829138755798)]

由此可知词嵌入模型可以有效地学习了物体间的关联关系。

此外可通过类比关系实验验证词嵌入模型对词与词之间关联关系建模的有效性进行分析，金庸的原著中杨过的主要武器为剑，基于这种关系，可以推测黄蓉的主要武器。当正例 positive=['杨过', '剑'], 负例 negative=['黄蓉']时输出的关联词为：

[('拳', 0.8988319635391235),
 ('招', 0.8886288404464722),
 ('棒', 0.876518726348877),
 ('拂尘', 0.8613254427909851),
 ('刀', 0.8608443737030029),
 ('剑法', 0.860496461391449),
 ('空手', 0.8603854179382324),
 ('快刀', 0.8594430088996887),

('掌', 0.8540631532669067),
('第二招', 0.8532459735870361)]
结果与黄蓉在小说中常用拳脚和洪七公传授的打狗棒法相符合。
利用 PCA 将 Word2Vec 模型学习得到的词向量投影到二维空间，可使用 matplotlib 库将词向量在二维平面上进行可视化。通过遍历 Word2Vec 模型的词汇表，将词向量保存到 rawWordVec 列表中，并建立词语到序号的映射关系。然后，利用 PCA 降维技术将词向量从高维空间降至二维。之后可绘制所有词向量的二维空间投影，并突出显示了一些特定词语的向量，如“剑”、“刀”、“郭靖”和“黄蓉”，以及它们在二维空间中的位置索引。结果如下图所示：

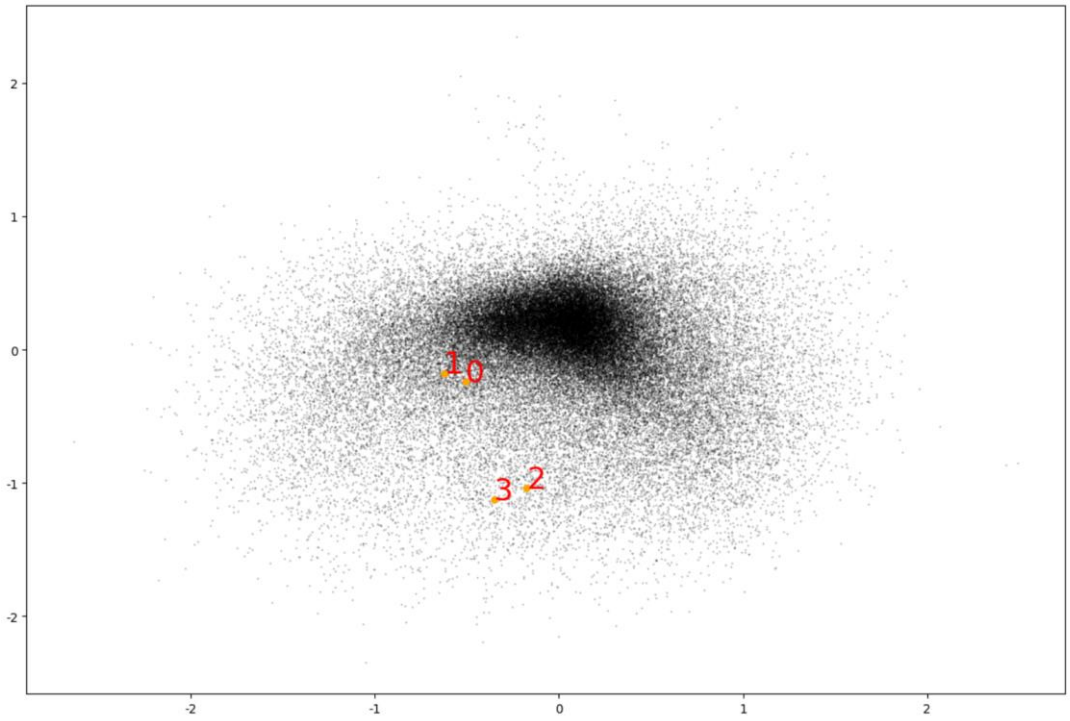

图 3 词向量二维空间投影

其中 0 和 1 分别为剑和刀，2 和 3 分别为郭靖和黄蓉。可以看到训练得到的词嵌入模型可以有效的表示近义词之间的距离关系，因而证明了其输出的词向量的有效性。

Conclusions

本研究探索了利用金庸武侠小说语料库训练词向量的有效性，并采用了 Word2Vec 神经网络模型进行一系列实验。通过实验结果的分析，得出以下结论：

Word2Vec 模型在金庸武侠小说语料库上表现出色：实验结果显示，基于金庸武侠小说语料库训练的 Word2Vec 模型能够生成高质量的词向量。这些词向量不仅在词汇语义相似性上表现出色，而且在词汇聚类和段落语义关联分析中也展示了良好的性能。这表明，利用金庸武侠小说这样的丰富中文文本进行词向量训练，可以为中文自然语言处理提供有力的工具，具有广泛的应用前景。

Word2Vec 模型在理解词汇语义关系上具有优秀表现：通过计算词向量间的语义距离和执行类比推理实验，验证了 Word2Vec 模型在捕捉词汇语义关系方面的能力。实验结果表明，

词嵌入模型可以有效地学习物体间的关联关系，并能够对词与词之间的关联关系进行准确建模。

词向量的可视化分析验证了模型的有效性：利用 PCA 将 Word2Vec 模型学习得到的词向量投影到二维空间，并通过 matplotlib 库进行可视化展示。结果显示，在二维空间中，训练得到的词向量能够有效地表示近义词之间的距离关系，进一步证明了 Word2Vec 模型输出的词向量的有效性。

References

[1] Zenchang Qin and Lao Wang (2023)，How to learn deep learning? Journal of Paper Writing, Vol. 3: 23: pp. 1-12.
[2] ….