NLP 第二次作业技术报告

基于 FNN/RNN/LSTM 的词向量学习与分析

2025-12-04

尹超
中国科学院大学
University of Chinese Academy of Sciences
2025
2313AI
Template:https://github.com/hzkonor/bubble-template
GitHub 地址:https://github.com/CarterYin/NLP_UCAS_2025

目录

I. Chapter 1 Introduction

本次作业内容如下：

1. (简答题)

   请下载调试FNN、RNN和LSTM模型的开源工具。利用北京大学标注的《人民日报》1998年1月份的分词语料，或者利用网络爬虫自己从互联网上收集足够多的英文文本语料，借助FNN或者RNN/LSTM开源工具，完成如下任务，并撰写一份实验报告：

   (1) 获得汉语或英语词语的词向量.

   (2) 对于同一批词汇，对比分别用FNN,RNN或LSTM获得的词向量的差异。

   (3) 利用你认为最好的词向量结果，对于随机选取的20个词汇分别计算与其词向量最相似的前10个单词，按相似度大小排序，人工对比排序结果是否与你的判断一致。

   (4) [选做]如果汉语和英语的词向量都学习到了，请对比同一个意思的汉语词汇和英语词汇，如“书”和‘book’，“工作”和‘work/ job’等，分析其向量距离。

说明

- 如果计算资源的限制，神经网络参数不必选择过大，例如：词表选择1000个左右单词即可,其余单词用代替；词向量的维度可设为10左右；神经网络的层数设置为1到2层；
- 可以使用某一种开放的深度学习框架，如TensorFlow或者PyTorch。
- 如果不借助开源工具和开放的深度学习框架，题目中的某些任务可以不做。

II. Chapter 2 Data Collection

本次作业使用的语料包括：

1. 中文语料：《人民日报》1998年1月份的分词语料，文件名为ChineseCorpus199801.txt，包含约100万词。
2. 英文语料：从互联网上爬取的英文文本语料，文件名为cd_snapshot_10MB.txt，包含约10MB文本数据。

III. Chapter 3 Model Implementation and Training

本次作业实现了三种不同的神经网络模型来学习词向量，分别是前馈神经网络（FNN）、循环神经网络（RNN）和长短期记忆网络（LSTM）。以下是各模型的简要介绍和训练过程。

III.1. 模型介绍

III.1.1. 前馈神经网络（FNN）

FNN模型通过固定大小的上下文窗口来预测目标词。对于每个目标词，模型使用其前后各三个词作为输入。模型结构包括嵌入层、隐藏层和输出层。训练过程中，使用交叉熵损失函数和Adam优化器进行优化。训练完成后，提取嵌入层的权重作为词向量。

III.1.2. 循环神经网络（RNN）

RNN 模型能够处理变长的输入序列，适合捕捉上下文信息。模型使用前五个词来预测下一个词。RNN 结构包括嵌入层、RNN 层和输出层。训练过程中，同样使用交叉熵损失函数和 Adam 优化器。训练完成后，提取嵌入层的权重作为词向量。

III.1.3. 长短期记忆网络（LSTM）

LSTM 模型是 RNN 的一种改进，能够更好地捕捉长期依赖关系。模型结构与 RNN 类似，但使用 LSTM 单元替代传统的 RNN 单元。训练过程与 RNN 相同，训练完成后提取嵌入层的权重作为词向量。

III.1.4. Training Details

所有模型均在相同的训练集上进行训练，使用相同的超参数设置。训练过程中监控损失值以防止过拟合。训练完成后，保存模型参数和词向量文件以供后续分析使用。

IV. Chapter 4 Results and Discussion

IV.1. 词向量模型对比 (FNN vs RNN vs LSTM)

本节对比了三种不同模型（FNN, RNN, LSTM）在中文和英文语料上训练得到的词向量效果。

IV.1.1. 中文语料分析

选取了“中国”、“发展”、“经济”等高频关键词，分析其在不同模型下的 Top-5 近义词。

Table 1 — 关键词“中国”的近义词对比

Rank	FNN	RNN	LSTM
1	大使馆 (0.321)	俄 (0.346)	沿 (0.345)
2	天天 (0.308)	途径 (0.329)	东方 (0.311)
3	中亚 (0.303)	研究所 (0.324)	傅 (0.310)
4	张家口 (0.294)	超越 (0.323)	大庆 (0.310)
5	广西 (0.294)	离开 (0.319)	判断 (0.292)

从定性分析来看，不同模型捕捉到的语义侧重点不同。FNN 倾向于共现频率高的实体（如“大使馆”），而 RNN 和 LSTM 捕捉到了更多地理或政治相关的语义（如“俄”、“东方”）。

IV.1.2. 英文语料分析

选取了“china”, “development”, “world”等关键词进行分析。

Table 2 — 关键词“china”的近义词对比

Rank	FNN	RNN	LSTM
1	germany (0.373)	instance (0.364)	africa (0.376)
2	suzhou (0.346)	hunan (0.352)	japan (0.354)
3	indonesia (0.341)	let (0.344)	fusion (0.323)
4	stakeholders (0.333)	canada (0.333)	country (0.316)
5	nation (0.315)	unchanged (0.317)	vietnam (0.316)

IV.1.3. 模型一致性分析

计算了不同模型之间 Top-10 近义词的 Jaccard 相似度。结果显示模型间的一致性较低（约 0.003），表明不同网络结构学习到的语义空间差异较大。

IV.2. 最佳模型深入分析 (LSTM)

基于 LSTM 模型的结果，对随机选取的 20 个词汇进行了人工评估。

IV.2.1. 中文词汇分析

Table 3 — 中文词汇人工评估示例

词汇	**Top-10** 相似词	人工分析
7 日	重要性, 5 日, 28, 揭示...	较好，捕捉到了日期特征
会议	揭晓, 增幅, 住房, 团拜会...	较好，关联了“团拜会”等会议类型
财政	文字, 使馆, 科委...	一般，关联性较弱

IV.2.2. 英文词汇分析

Table 4 — 英文词汇人工评估示例

Word	Top-10 Neighbors	Analysis
research	robotics, india, academy...	Good, related to academic topics
years	mean, porcelain, decade...	Good, “decade” is a synonym
asian	suzhou, central, african...	Good, captures geographic context

总体而言，模型在具体名词和数字上的聚类效果优于抽象词和功能词。

IV.3. 跨语言词向量对比

对比了同一语义在中文和英文模型中的向量距离。

Table 5 — 跨语言词向量距离对比

中文	英文	余弦相似度	欧氏距离
书	book	0.0419	14.72
中国	china	−0.0411	15.51
合作	cooperation	0.1350	13.73
世界	world	−0.1464	16.57

结果显示，所有词对的余弦相似度均接近于0。这是因为中英文词向量是在两个独立的向量空间中训练的，缺乏跨语言对齐（Alignment），因此直接比较向量数值没有物理意义。

V. Chapter 5 Conclusion

本次实验通过 FNN、RNN 和 LSTM 三种模型，在中文（人民日报 1998）和英文（Web Snapshot）语料上完成了词向量的训练与分析。主要结论如下：

1. 模型性能差异：三种模型在捕捉语义特征上表现出显著差异。虽然都能在一定程度上聚类相关词汇，但模型间的一致性较低。LSTM 模型在处理长距离依赖和具体实体名词（如地名、数字、特定领域术语）时表现出相对较好的效果。
2. 语义捕捉局限性：受限于训练语料规模（约 10MB）和训练时间，生成的词向量质量总体一般。对于具体名词的聚类效果优于抽象概念和功能词，且近义词列表中存在一定的噪声。
3. 跨语言独立性：实验验证了不同语言模型训练出的向量空间是相互独立的。在未进行对齐操作的情况下，中英文对应词汇的向量在空间中几乎正交，无法直接通过距离度量进行语义对比。

V.1. 遇到的问题与解决方案

在实验过程中，主要遇到了以下问题，并提出了相应的改进思路：

1. 词向量语义质量不高
   - 问题描述：部分高频词的近义词列表包含大量无关词汇，语义聚合度低。
   - 原因分析：训练语料过小（仅 10MB），导致模型无法充分学习词汇的共现模式；同时为了节省时间，训练轮数较少，模型可能处于欠拟合状态。
   - 解决方案：
     - 扩大语料库规模至 GB 级别（如使用维基百科 Dump）。
     - 增加训练轮数（Epochs）和调整超参数（如词向量维度、窗口大小）。
     - 引入预训练模型（如 Word2Vec, GloVe）或上下文相关模型（BERT）以提升质量。
2. 跨语言向量无法直接比较
   - 问题描述：直接计算中英文对应词（如“书”和“book”）的余弦相似度，结果接近0。
   - 原因分析：不同语言的向量空间是随机初始化并独立演化的，缺乏统一的坐标系。

- 解决方案：
  - 使用跨语言词向量对齐技术（Cross-lingual Word Embedding Alignment）。
  - 利用少量种子词典训练一个线性映射矩阵，将一个语言的向量空间映射到另一个语言的空间中。

3. 未登录词（**OOV**）处理
   - 问题描述：测试集中出现训练集中未包含的词汇，导致无法计算向量。
   - 解决方案：
     - 使用 `<UNK>` 标记统一处理低频词。
     - 采用基于子词（Subword）的模型（如 FastText），利用字符级 n-gram 信息生成未登录词的向量。

VI. Appendix: Code Listings

代码仓库地址：https://github.com/CarterYin/NLP_UCAS_2025