Report of Deep Learning for Natural Language Processing III

Yucheng Wang

wang_eil@126.com

Abstract

以金庸的 16 篇小说作为语料库，基于 Word2Vec 模型训练得到词向量，通过 t-SNE（t-Distributed Stochastic Neighbor Embedding）来将前 100 个高维的 Word2Vec 词向量降维至二维进行可视化，观察词向量结果，通过搜索相似词和比较段落语义关联程度来观察词向量的有效性。

Introduction

Word2Vec 是一种将词汇转换为向量表示（词向量）的模型，核心思想是通过神经网络模型将词汇映射到一个低维向量空间中，使得在该空间中，具有相似语义的词汇距离较近。Word2Vec 模型在自然语言处理（NLP）领域中广泛应用，尤其是在文本分类、情感分析、机器翻译等任务中表现优异。

Word2Vec 包含了两种训练方法：

① CBOW（Continuous Bag of Words）：

通过预测上下文词（context words）来预测目标词（target word）。给定一个词的上下文，模型试图预测这个词是什么。举例来说，对于句子 "The cat sits on the mat"，模型会用 "The", "cat", "on", "the", "mat" 这些词来预测 "sits"。

② Skip-gram：

与 CBOW 相反，使用目标词来预测上下文词。给定一个词，模型试图预测它的上下文词。例如，对于同一句子 "The cat sits on the mat"，Skip-gram 模型会用 "sits" 这个词来预测 "The", "cat", "on", "the", "mat"。

本文使用 CBOW 方法，再通过 t-SNE 将前 100 词降维可视化，观察词向量的有效性。相同类别的词向量在图上会比较接近。通过搜索相似词和比较段落语义关联程度来观察词向量的有效性。

Methodology

首先使用 jieba 分词对 16 篇小说进行分词，作为 Word2Vec 模型的语料库。在分词前进行删除隐藏符号和非中文字符。

Jieba 库是一款优秀的用于中文分词的库，它利用一个中文词库，确定汉字之间的关联概率，汉字间概率大的组成词组，形成分词结果。Jieba 库支持精确、全模式、搜索引擎三种分词模式。本文中对经过预处理的文本使用 Jieba 精确模式，它将一段文本切分为若干个中文单词，不会产生冗余，是最适合词频统计的模式。

然后使用分词后的语料训练 Word2Vec 模型，通过调整词向量维度、上下文窗口大小观察其对分词结果的影响。通过观察降维后的词向量图，并通过 similar_words 方法搜索最接近的词向量，通过 cosine_similarity 方法计算段落之间的相似度，从而判断词向量的有效性。

Experimental Studies

迭代次数为 200，词向量维度在 200 和 300 中选择，窗口大小在 5 和 7 中选择，共进行四组实验，分组如下：

窗口大小 / 词向量维度	5	7
200	模型 1	模型 2
300	模型 3	模型 4

训练得到了四个模型：

模型 1：word2vec_dim200_win5.model；

模型 2：word2vec_dim200_win7.model；

模型 3：word2vec_dim300_win5.model；

模型 4：word2vec_dim300_win7.model。

通过 t-SNE 进行降维可视化，结果如下：

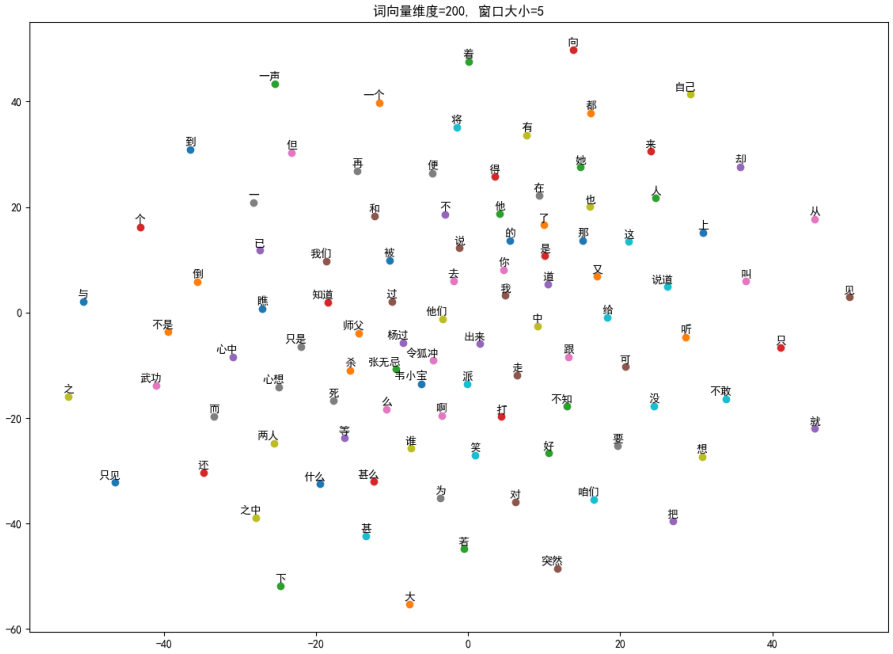

图 1 词向量维度=200，窗口大小=5 的降维结果

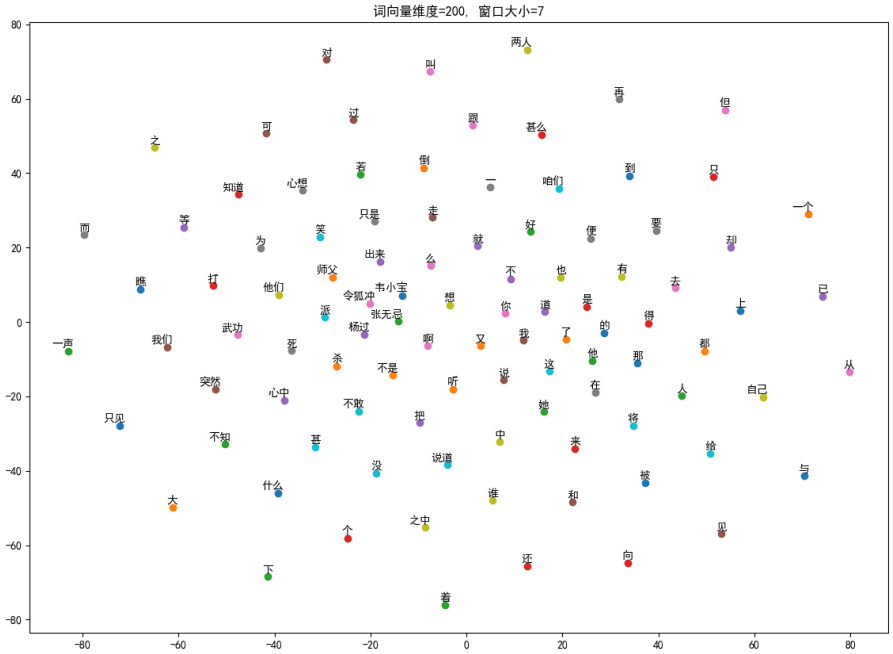

图 2 词向量维度=200，窗口大小=7 的降维结果

、

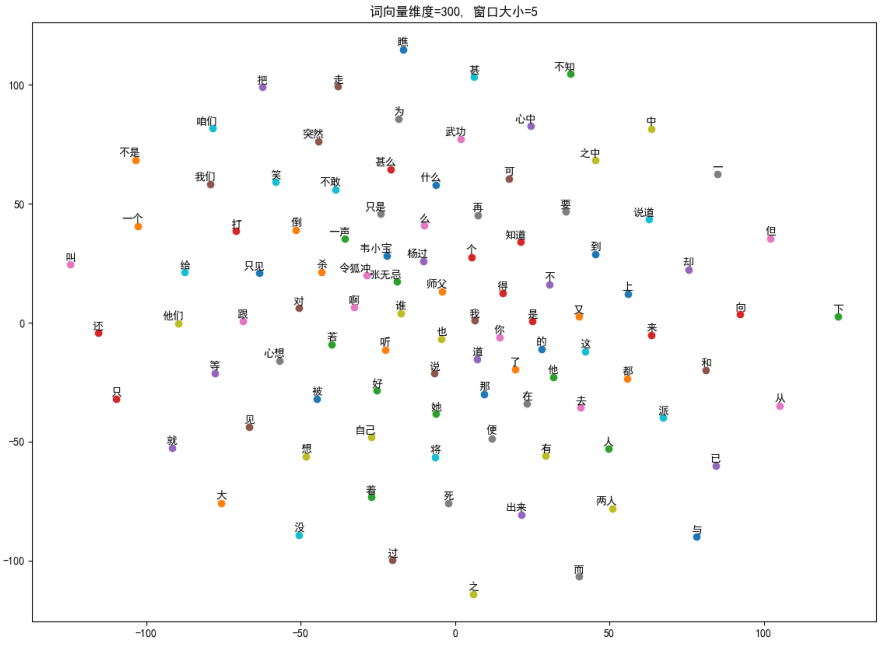

图 3 词向量维度=300，窗口大小=5 的降维结果

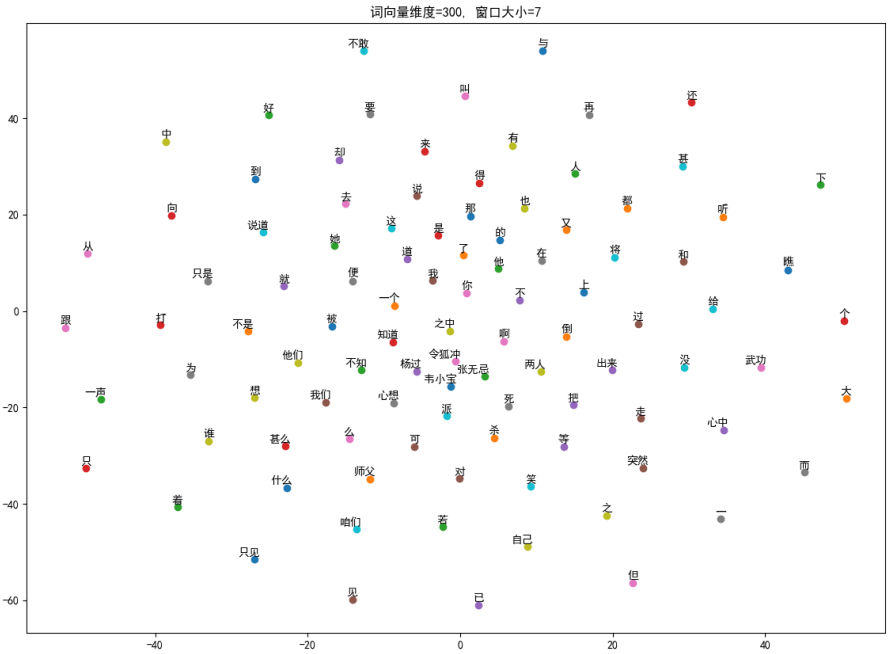

图 4 词向量维度=300，窗口大小=7 的降维结果

通过 similar_words 搜索 “武功”的结果如下：

	模型 1	模型 2	模型 3	模型 4
武功	功夫、剑术 武艺、本领	武艺、本领、 剑术、胜	武艺、功夫、 剑术、本领	功夫、本领、 玩艺儿、武艺

通过 cosine_similarity 对以下三个段落比较使用 word2vec_dim200_win5.model 模型生成的词向量。

段落为：

"一转头，只见地下明晃晃的撒著十几枚银针，针身镂刻花纹，打造得极是精致。他俯身一枚枚的拾起，握在左掌，忽见银针旁一条大蜈蚣肚腹翻转，死在地下。他觉得有趣，低头细看，见地下蚂蚁死了不少，数步外尚有许多蚂蚁正在爬行。他拿一枚银针去拨弄几下，那几只蚂蚁兜了几个圈子，便即翻身僵毙，连试几只小虫都是如此。"来自金庸小说《神雕侠侣》。

"天灵星一抬头，和古浊飘那锐利的目光撞个正着，他心中一动，升起一个念头，猛的走前两步，一把拍向古浊飘的肩头，笑道：“一掷千金无吝色，神州谁是真豪杰，公子的确是快人。”古浊飘眼神一动，已觉一股极强的力道压了下来，暗忖道：“这老儿倒是个内家高手。”随即微微一笑，在这力道尚未使满之际，伸出手去，像是去拉天灵星的膀子，口中却笑道：“孙老英雄过奖了。”"来自古龙小说《残金缺玉》。

"2014 年乌克兰危机开启了俄罗斯与乌克兰双边关系恶化的进程，尤其是克里米亚全民公投并入俄罗斯之后使得两国关系长期处于敌对状态，再加上乌东地区的分离倾向以及俄罗斯对乌东两个“共和国”的支持，加深了俄乌边境军事对抗局面，这些都是今日俄乌冲突爆发的前奏与预演。"来自光明日报新闻。

得到的相似度矩阵如下：

[[0.99999976 0.86264086 0.75831854]
 [0.86264086 1.0000001   0.7522998]
 [0.75831854 0.7522998   1.0000001]]

Conclusion

通过 Word2Vec 建立的词向量模型验证了词向量的有效性，对“武功”的搜索结果与认知上相同。经过可视化之后，发现在每个图中，人名都会在中间聚集。这可能说明，词出现的个数也会影响词向量的结果。

在实验中，进行了四组对照试验，在搜索“武功”时，其结果并没有显著区别。但是窗口增大时，对应的概率由最高 0.42 降低至最高 0.39，这可能说明较大的窗口使得模型考虑了更多的上下文信息之后，词语语义更加多元。

在进行段落相似度对比时，可以发现，金庸与古龙小说同为武侠小说，相似度更高，两者与新闻的相似度都较低。