Report of Deep Learning for Natural Langauge Processing

Tuo Zhao
975352925@qq.com

Abstract

本报告对包含 16 部小说的中文语料库进行数据预处理作为数据集，利用 Word2Vec 模型训练词向量，通过计算词向量之间的语意距离、某一类词语的聚类、某些段落直接的语意关联来验证所训练的词向量的有效性。

Introduction

Word2Vec 是一种经典的自然语言处理技术，它的目标是将单词转换为密集向量，使得具有相似语义的单词在向量空间中彼此靠近。这种技术是由 Google 的 Tomas Mikolov 等人在 2013 年提出的，它基于两种主要的模型：连续词袋模型（CBOW）和 Skip-gram 模型。

在连续词袋模型中，模型尝试根据上下文来预测目标单词，而在 Skip-gram 模型中，模型尝试根据目标单词来预测上下文。这两种方法都使用了神经网络，通常是浅层的前馈神经网络，通过学习大量文本语料库，模型可以逐渐学习到单词之间的语义关系。

Word2vec 生成的向量空间具有一些有趣的性质，例如，在这个空间中，通过简单的向量运算，可以进行诸如词语之间的类比推理，比如 "king - man + woman = queen"。这种能力使得 Word2vec 成为许多自然语言处理任务的重要工具，如词义相似度计算、文档分类、情感分析等等。ultimate Word2vec 为自然语言处理领域带来了深远的影响，促进了语义理解的发展。此外，Word2vec 的思想也启发了许多其他词嵌入技术的发展，如 GloVe、FastText 等

本文利用 Word2Vec 模型训练词向量，通过计算词向量之间的语意距离、某一类词语的聚类、某些段落直接的语意关联来验证所训练的词向量的有效性。

Methodology

具体来说，分类流程有以下步骤：

1. 准备语料库：选择一个中文文本作为实验的语料库。本报告中，将 16 部金庸小说进行合并得到完整文本。
2. 文本预处理：对语料库进行预处理。删除所有的隐藏符号、非中文字符和标点符号。对文本进行分词并过滤停用词
3. 训练模型：使用gensim库的Word2Vec模型对处理后的语料库进行训练。
4. 保存和加载Word2Vec模型：将训练好的模型保存到文件，并在需要时加载模型。
5. 词语相似度计算：计算两个词语之间的相似度得分与语意距离。
6. KMeans聚类与t-SNE降维与可视化：对词向量进行KMeans聚类，并计算聚类的轮廓系数。使用t-SNE对指定簇的词向量进行降维，并绘制散点图。

Experiment Results

Task1：计算词向量之间的语意距离

本次实验对比词设置如下：（武林、江湖）、（刀光、剑影）、（马匹、飞鸟）、（杨过、小龙女）。在词向量维度为 100、训练轮次为 50 的条件下，进行相似度实验。相似度越接近 1，说明词向量对相似度越高，实验结果如下表：

词向量对	武林/江湖	刀光/剑影	马匹/飞鸟	杨过/小龙女
语意距离	0.707	0.847	0.558	0.828

表 1 不同词向量对的余弦相似度

从表 1 中可以看出，（武林、江湖）、（刀光、剑影）、（杨过、小龙女）的相似度较高，而（马匹、飞鸟）的相似度较低。分析原因可能如下：

刀光和剑影的相似度高，而马匹和飞鸟的相似度低可能是由于以下原因：

1.语境差异：刀光和剑影在语境上往往与战斗、武侠等相关联，因此它们的语义可能在某种程度上相似。相比之下，马匹和飞鸟在语境上可能更加分散，它们代表了不同的生物类别和行为，因此语义之间的关联性可能较低。

2.语言习惯：在文学作品或者常用语中，刀光和剑影常常被放在一起描述，因为它们经常同时出现在同一场景中。相比之下，马匹和飞鸟的搭配可能较少，因此它们的语义关联性可能较低。

杨过和小龙女的相似度高可能有以下原因：

1. 共现频率高：在金庸的作品中，杨过和小龙女是一个著名的情侣角色，他们经常一起出现在小说情节中。因此，根据共现频率高的原则，这两个词在文本中的相似度可能会较高。

2. 语境关联：杨过和小龙女的故事情节通常涉及到爱情、江湖义气等主题，这些共同的语境会增强它们之间的语义关联性，使得它们在词向量空间中的距离更近。

Task2：对某一类词语进行聚类，检验词向量

本次实验使用 K-Means 聚类算法对词向量进行聚类。

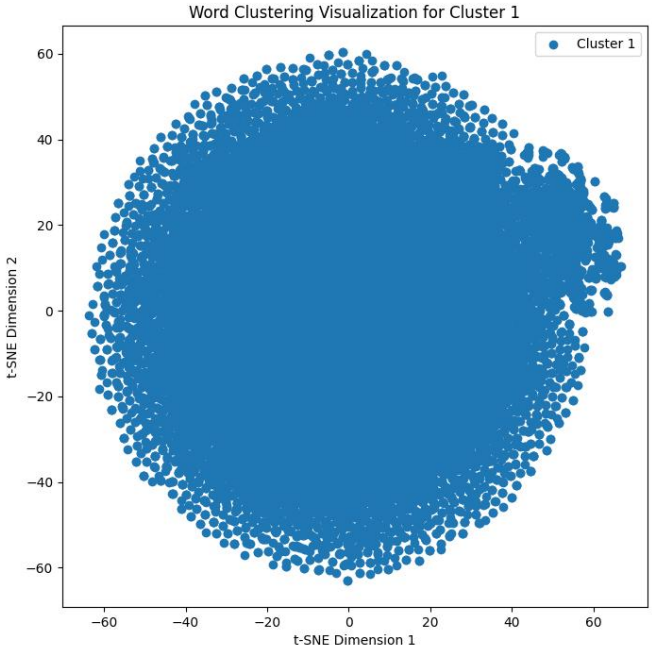

图 1 对索引为 1 的簇进行聚类绘制

取该簇部分词汇展示：

令狐冲, 张无忌, 教主, 陈家洛, 石破天, 虚竹, 欧阳锋, 萧峰, 洪七公, 岳不群, 张翠山, 太后, 吴三桂, 谢逊, 岳灵珊, 青青, 王语嫣, 慕容复, 乔峰, 周芷若, 恒山, 程灵素, 赵敏, 法王, 段正淳, 木婉清, 田伯光, 华山派, 徐天宏, 仪琳, 魔教, 罗刹, 文泰来, 王爷, 姊姊, 明教, 南海, 鸠摩智, 殷素素, 海老公, 向问天, 任我行, 游坦之, 姑姑, 欧阳克, 喇嘛, 梅超风, 赵志敬, 契丹, 左冷禅, 逍, 张三丰, 星宿, 丁当。

计算轮廓系数为 0.74576735496521，轮廓系数越接近 1，聚类越准确。

从图 1 中可以看出在 t-SNE 降维后的空间中，单一簇的数据点呈现出球形的分布。表明数据点在高维空间中均匀地分布在球形区域内，没有明显的聚集或分散趋势。这可能意味着数据点的密度分布比较均匀，没有明显的聚类结构或异常值。数据点之间的相似性较高，没有明显的子群体或亚群体。

从聚类后的词汇组可以看出这些词语与金庸武侠小说中的人物、地点、组织等相关，说明聚类较准确。

Task3：计算段落的语意关联

计算段落之间的语意关联的步骤如下：

1.段落向量化：首先，将每个段落表示为一个向量。可以使用类似Word2Vec的技术，将段落中的每个词向量化，并取平均值或加权平均值作为段落向量。

2.计算段落间相似度：使用向量表示的段落，可以通过计算它们之间的相似度来衡量它们的语义关联性。常用的方法包括余弦相似度、欧氏距离、曼哈顿距离等。

3.相似度阈值：根据实际需求，可以设定一个相似度阈值，超过这个阈值的段落视为语义相关，否则视为不相关。

本次实验段落选择如下：

第一个段落："颜烈跨出房门，只见过道中一个　　颜烈跨出房门，只见过道中一个中年士人拖着鞋皮，踢踏踢踏的直响，一路打着哈欠迎面过来。那士人似笑非笑，挤眉弄眼，一副惫懒神气，全身油腻，衣冠不整，满脸污垢，看来少说也有十多天没洗脸了，拿着一柄破烂的油纸黑扇，边摇边行。颜烈见这人衣着明明是个斯文士子，却如此肮脏，不禁皱了眉头，加快脚步，只怕沾到了那人身上的污秽。突听那人干笑数声，声音甚是刺耳，经过他身旁时，顺手伸出折扇，在他肩头一拍。颜烈身有武功，这一下竟没避开，不禁大怒，喝道："干什么？"那人又是一阵干笑，踢踏踢踏的向前去了，只听他走到过道尽头，对店小二道："喂，伙计啊，你别瞧大爷身上破破烂烂的，大爷可有的是银子。有些小子可邪着哪，他就是仗着身上光鲜吓人。招摇撞骗，勾引妇女，吃白食，住白店，全是这种小子，你得多留点儿神。稳稳当当的，让他先交了房饭钱再说。"也不等那店小二答腔，又是踢踏踢踏的走了。颜烈更是心头火起，心想好小子，这话不是冲着我来么？店小二听那人一说，斜眼向他看了一眼，不禁起疑，走到他跟前，哈了哈腰，陪笑道："您老别见怪，不是小的无礼……"颜烈知他意思，哼了一声道："把这银子给存在柜上！"伸手往怀里一摸，不禁呆了。他囊里本来放着四五十两银子，一探手，竟已空空如也。店小二见他脸色尴尬，只道穷酸的话不错，神色登时不如适才恭谨，挺腰凸肚的道："怎么？没带钱么？"颜烈道："你等一下，我回房去拿。"他只道匆匆出房，忘拿银两，那知回入房中打开包裹一看，包里几十两金银竟然尽皆不翼而飞。这批金银如何失去，自己竟是茫然不觉，那倒奇了，寻思："适才包氏娘子出去解手，我也去了茅房一阵，前后不到一柱香时分，怎地便有人进房来做了手脚？嘉兴府的飞贼倒是厉害。"店小二在房门口探头探脑的张望，见他银子拿不出来，发作道："这女娘是你原配妻子吗？要是拐带人口，可要连累我们呢！"包惜弱又羞又急，满脸通红。颜烈一个箭步纵到门口，反手一掌，只打得店小二满脸是血，还打落了几枚牙齿。店小二捧住脸大嚷大叫："好哇！住店不给钱，还打人哪！"颜烈在他屁股上加了一脚，店小二一个筋斗翻了出去。包惜弱惊道："咱们快走吧，不住这店了。"颜烈笑道："别怕，没了银子问他们拿。"端

了一张椅子坐在房门口头。过不多时，店小二领了十多名泼皮，抡棒使棍，冲进院子来。颜烈哈哈大笑，喝道：“你们想打架？”忽地跃出，顺手抢过一根杆棒，指东打西，转眼间打倒了四五个，那些泼皮平素只靠逞凶使狠，欺压良善，这时见势头不对，都抛下棍棒，一窝蜂的挤出院门，躺在地下的连爬带滚，唯恐落后。包惜弱早已吓的脸上全无血色，颤声道：“事情闹大了，只怕惊动了官府。”颜烈笑道：“我正要官府来。”包惜弱不知他的用意，只得不言语了。过不半个时辰，外面人声喧哗，十多名衙役手持铁尺单刀，闯进院子，把铁链抖的当啷当啷乱响，乱嘈嘈的叫道：“拐卖人口，还要行凶，这还了得？凶犯在那里？”颜烈端坐椅上不动。众衙役见他衣饰华贵，神态俨然，倒也不敢贸然上前。带头的捕快喝道：“喂，你叫什么名字？到嘉兴府来干什么？”颜烈道：“你去叫盖运聪来!”盖运聪是嘉兴府的知府，众衙役听他直斥上司的名字，都是又惊又恐。那捕快道：“你失心疯了么？乱呼乱叫盖大爷。”颜烈从怀里取出一封信来，往桌上一掷，抬头瞧着屋顶，说道：“你拿去给盖运聪瞧瞧，看他来是不来？”那捕快取信件，见了封皮上的，吃了一惊，但不知真伪，低声对众衙役道：“看着他，别让他跑了。”随即飞奔而出。包惜弱坐在房中，心里怦怦乱跳，不知吉凶。"

第二个段落："完颜洪烈眼前一花，只见一个道人手中托了一口极大的铜缸，迈步走上楼来，定睛看时，只吓得心中突突乱跳，原来这道人正是长春子丘处机。完颜洪烈这次奉父皇之命出使宋廷，要乘机阴结宋朝大官，以备日后入侵时作为内应。陪他从燕京南来的宋朝使臣王道乾趋炎附势，贪图重贿，已暗中投靠金国，到临安后替他拉拢奔走。那知王道乾突然被一个道人杀死，连心肝首级都不知去向。完颜洪烈大惊之余，生怕自己阴谋已被这道人查觉，当即带同亲随，由临安府的捕快衙役领路，亲自追拿刺客。追到牛家村时与丘处机遭遇，不料这道人武功高极，完颜洪烈尚未出手，就被他一甩手箭打中肩头，所带来的衙役随从被他杀的干干净净。完颜洪烈如不是在混战中先行逃开，又得包惜弱相救，堂堂金国王子就此不明不白的葬身在这小村之中了。完颜洪烈定了定神，见他目光只在自己脸上掠过，便全神贯注的瞧着焦木和那七人，显然并未认出自己，料想那日自己刚探身出来，便给他羽箭掷中摔倒，并未看清楚自己面目，当即宽心，再看他手中托的那口大铜缸时，一惊之下，不由得欠身离椅。"

段落1与段落2的语义相似度：**0.8942413330078125**，说明段落之间的相似度高，模型训练的词向量效果好。

Conclusion

本次实验结果表明，WordVec 模型有效训练中文语料库的词向量，在三项任务中表现良好。

Reference

[1]https://paddlepedia.readthedocs.io/en/latest/tutorials/sequence_model/word_representation/word2vec.html