Report of Deep Learning for Natural Language Processing

种雨萌

2775965921@qq.com

Abstract

本篇报告从金庸的小说中选取部分经典作品作为语料库，利用 Word2Vec 模型来训练词向量，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三种方法，验证了词向量的有效性。

Introduction

传统的自然语言处理将词看作是一个个孤立的符号，这样的处理方式对于系统处理不同的词语没有提供有用的信息。词映射(word embedding)实现了将一个不可量化的单词映射到一个实数向量。Word embedding 能够表示出文档中单词的语义和与其他单词的相似性等关系，被广泛应用于推荐系统和文本分类中。Word2Vec 模型是 Word embedding 中广泛应用的模型。Word2Vec 使用一层神经网络将 one-hot（独热编码）形式的词向量映射到分布式形式的词向量，同时使用 Hierarchical softmax，negative sampling 等技巧进行训练速度上的优化。

本次实验将利用给定语料库，利用 1～2 种神经语言模型（基于 Word2Vec，LSTM，GloVe 等模型）来训练词向量，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、或者其他方法来验证词向量的有效性。

Methodology

M1：语料处理

在读取语料前，需要对文本进行预处理。首先利用 jieba 分词对语料进行分词，去掉文本中一些无意义的词语和标点符号等，并将处理后的语料重新保存进新的文件夹中。

M2：模型训练

词映射(word embedding)实现了将一个不可量化的单词映射到一个实数向量。

Word embedding 能够表示出文档中单词的语义和与其他单词的相似性等关系。它已经被广泛应用在了推荐系统和文本分类中。Word2Vec 模型则是 Word embedding 中广泛应用的模型。Word2Vec 使用一层神经网络将 one-hot（独热编码）形式的词向量映射到分布式形式的词向量。使用了 Hierarchical softmax，negative sampling 等技巧进行训练速度上的优化。

Word2Vec 是从大量文本语料中以无监督的方式学习语义知识的一种模型，它通过学习文本来用词向量的方式表征词的语义信息，即通过一个嵌入空间使得语义上相似的单词在该空间内距离很近。例如，man 和 woman 属于语义上很相近的词，当这两个词映射到新的空间后，得到的嵌入向量就很接近。

Word2Vec 有两种模型——CBOW(Continuous Bag-of-Word)和 SkipGram。这两者的区别在于，CBOW 模型是以上下文词汇预测当前词，即用w_{t-2}，w_{t-1}，w_{t+1}和w_{t+2}去预测w_t；而 SkipGram 模型是以当前词预测其上下文词汇，即用w_t去预测w_{t-2}，w_{t-1}，w_{t+1}和w_{t+2}。本次实验采用的是 CBOW 模型，其对词向量训练步骤包括：

（1）将上下文词进行 one-hot 表征作为模型的输入，其中词汇表的维度为 V，上下文单词数量为 C；

（2）将所有上下文词汇的 one-hot 向量分别乘以输入层到隐层的权重矩阵 W；

（3）将上一步得到的各个向量相加取平均作为隐藏层向量；

（4）将隐藏层向量乘以隐藏层到输出层的权重矩阵 W’；

（5）将计算得到的向量做 softmax 激活处理得到 V 维的概率分布，取概率最大的索引作为预测的目标词。

M3：有效性验证

模型训练完毕之后，为验证词向量的有效性，选择以下三种方法进行实验。

（1）给定一个词向量，计算其他词向量与目标词的语义距离，输出语义距离最高的 10 个词；

（2）选择一组给定词，对其进行快速聚类，类别数设置为 2；

（3）任意选取两个段落，计算段落之间的语义关联。

Experimental Studies

本实验针对金庸先生的小说作品，利用 Word2Vec 模型进行词向量训练，并通过三种方法分别进行词向量的有效性验证。

1、《白马啸西风》的训练结果如下表所示：

	苏普	**关联度**
方法一： 词向量语义距离	阿曼	0.3799048662185669
	敌人	0.2615184187889099
	似乎	0.22161105275154114
	渐渐	0.18799838423728943
	一下	0.1792408972978592
	车尔库	0.17220322787761688
	苏鲁克	0.16633810102939606
	朋友	0.1600337028503418
	狼皮	0.15960870683193207
	能	0.15952466428279877
方法二： 词语聚类	**词组**	**类别**
	李文秀	0
	苏普	0
	高昌	1
	阿曼	0
方法三： 段落语义关联	**所选段落关联度**	
	0.1667167693376541	

注：所选段落都提及华辉等。

2、《碧血剑》的训练结果如下表所示：

	袁承志	**关联度**
方法一： 词向量语义距离	青青	0.4585103988647461
	师父	0.3058967888355255
	何铁手	0.3041801452636719

	人	0.300434410572052
	穆人清	0.2975253760814667
	又	0.2962271571159363
	洪胜海	0.291443407535553
	曹化淳	0.27861058712005615
	崔希敏	0.274262011051178
	温青	0.26898902654647827
方法二： 词语聚类	**词组**	**类别**
	袁承志	0
	青青	0
	华山	1
方法三： 段落语义关联	**所选段落关联度**	
	0.4186725616455078	

注：所选段落都提及焦公礼、袁承志、青青等人。

3、《飞狐外传》的训练结果如下表所示：

方法一： 词向量语义距离	**胡斐**	**关联度**
	王剑英	0.3115085959434509
	袁紫衣	0.27934515476226807
	田归农	0.2561693787574768
	马春花	0.24164317548274994
	桑飞虹	0.23766213655471802
	穴	0.23481272161006927
	福康安	0.2281418889760971
	宗雄	0.22582383453845978
	但	0.22461870312690735
	敌人	0.21858154237270355
方法二： 词语聚类	**词组**	**类别**
	胡斐	0
	袁紫衣	0

	凤天南	1
	程灵素	0
方法三： 段落语义关联	**所选段落关联度**	
	0.12044910341501236	

注：所选段落都提及胡斐、苗人凤等人。

4、《连城诀》的训练结果如下表所示：

方法一： 词向量语义距离	**狄云**	**关联度**
	丁大哥	0.28601959347724915
	吴坎	0.25613757967948914
	师妹	0.2288082391023636
	戚长发	0.20137614011764526
	丁	0.1930244117975235
	一下	0.18922361731529236
	周圻	0.18836858868598938
	头儿	0.186028391122818
	水岱	0.1855621337890625
	避开	0.18164145946502686
方法二： 词语聚类	**词组**	**类别**
	狄云	0
	戚芳	0
	神照经	1
方法三： 段落语义关联	**所选段落关联度**	
	0.13630717992782593	

注：所选段落都提及狄云、宝象等人。

5、《鹿鼎记》的训练结果如下表所示：

方法一： 词向量语义距离	**韦小宝**	**关联度**
	康熙	0.5279734134674072
	太后	0.44994306564331055

	双儿	0.40784671902656555
	海老公	0.39239194989204407
	皇上	0.39149200916290283
	多隆	0.37391039729118347
	陈近南	0.37029194831848145
	小桂子	0.3686050772666931
	吴之荣	0.3633728623390198
	公主	0.3464660346508026
方法二： 词语聚类	**词组**	**类别**
	韦小宝	0
	天地会	1
	康熙	0
方法三： 段落语义关联	**所选段落关联度**	
	0.5860589146614075	

注：所选段落都提及阿珂、韦小宝等人。

6、《射雕英雄传》的训练结果如下表所示：

方法一： 词向量语义距离	**郭靖**	**关联度**
	黄蓉	0.5580871105194092
	完颜康	0.4050754904747009
	洪七公	0.3840724229812622
	裘千仞	0.37525075674057007
	欧阳锋	0.37056398391723633
	周伯通	0.350513756275177
	欧阳克	0.3469294011592865
	柯镇恶	0.34149718284606934
	穆念慈	0.3322194218635559
	华筝	0.32975083589553833
方法二： 词语聚类	**词组**	**类别**
	郭靖	0

	黄蓉	0
	杨康	1
方法三： 段落语义关联	**所选段落关联度**	
	0.3608575463294983	

注：所选段落都提及欧阳锋、黄药师等人。

7、《神雕侠侣》的训练结果如下表所示：

	杨过	**关联度**
方法一： 词向量语义距离	小龙女	0.5851811170578003
	周伯通	0.48364973068237305
	李莫愁	0.4601018726825714
	尹志平	0.4270373582839966
	裘千尺	0.39618033170700073
	陆无双	0.38340988755226135
	绿萼	0.38210374116897583
	洪凌波	0.37160399556159973
	过儿	0.3692546486854553
	武修文	0.3669302463531494
方法二： 词语聚类	**词组**	**类别**
	杨过	0
	小龙女	0
	李莫愁	1
方法三： 段落语义关联	**所选段落关联度**	
	0.08093375712633133	

注：所选段落均提及李莫愁等。

8、《书剑恩仇录》的训练结果如下表所示：

	陈家洛	**关联度**
方法一： 词向量语义距离	乾隆	0.4253509044647217
	霍青桐	0.40458571910858154
	徐天宏	0.40095406770706177

	香香公主	0.3742794692516327
	陆菲青	0.35084494948387146
	陈正德	0.3361518979072571
	李沅芷	0.32526612281799316
	文泰来	0.32245659828186035
	张召重	0.3119031488895416
	袁士霄	0.303833544254303
方法二： 词语聚类	**词组**	**类别**
	陈家洛	0
	红花会	1
	乾隆	0
方法三： 段落语义关联	**所选段落关联度**	
	0.4061174690723419	

注：所选段落均提及狼群等。

9、《天龙八部》的训练结果如下表所示：

方法一： 词向量语义距离	**段誉**	**关联度**
	王语嫣	0.3730698525905609
	虚竹	0.3114943206310272
	乔峰	0.28636467456817627
	乔大爷	0.28016695380210876
	段郎	0.2755787670612335
	段正淳	0.2731360197067261
	慕容复	0.27114152908325195
	段誉心	0.2648598849773407
	钟灵	0.26202207803726196
	王夫人	0.257327675819397
方法二： 词语聚类	**词组**	**类别**
	段誉	0
	萧峰	0

	阿紫	1
方法三： 段落语义关联	**所选段落关联度**	
	0.3561622202396393	

注：所选段落均提及萧峰、阿紫等人。

10、《侠客行》的训练结果如下表所示：

方法一： 词向量语义距离	**石破天**	**关联度**
	丁不四	0.33456388115882874
	丁当	0.2962532043457031
	就此	0.2650655508041382
	便	0.25981253385543823
	小丐	0.21749667823314667
	几	0.20761412382125854
	胆子	0.2055637389421463
	有	0.20470359921455383
	还是	0.20055967569351196
	说	0.19981548190116882
方法二： 词语聚类	**词组**	**类别**
	石破天	0
	阿绣	1
	石中玉	0
方法三： 段落语义关联	**所选段落关联度**	
	0.17039291560649872	

注：所选段落均提及阿绣、石破天等。

11、《笑傲江湖》的训练结果如下表所示：

方法一： 词向量语义距离	**令狐冲**	**关联度**
	岳不群	0.4195218086242676
	林平之	0.38485005497932434
	令狐冲笑	0.34940361976623535
	田伯光	0.3465957045555115

	盈盈	0.3223159611225128
	便	0.31715625524520874
	劳德诺	0.3059697449207306
	令狐大哥	0.2992016077041626
	婆婆	0.29781046509742737
	岳夫人	0.29257866740226746
方法二： 词语聚类	**词组**	**类别**
	令狐冲	0
	岳不群	0
	东方不败	1
方法三： 段落语义关联	**所选段落关联度**	
	0.2478450983762741	

注：所选段落均提及令狐冲、仪清等。

12、《雪山飞狐》的训练结果如下表所示：

方法一： 词向量语义距离	**胡一刀**	**关联度**
	金面佛	0.322707861661911
	胡兄	0.2446996420621872
	帐	0.19947610795497894
	夫人道	0.1977846622467041
	亲手	0.19081337749958038
	手里	0.1821662038564682
	酒	0.1708204299211502
	从来	0.166363924741745
	实	0.16412031650543213
	喝酒	0.1640172153711319
方法二： 词语聚类	**词组**	**类别**
	胡斐	1
	胡一刀	0
	苗人凤	0

方法三： 段落语义关联	**所选段落关联度**
	0.6081887483596802

注：胡一刀和苗人凤是一辈人；所选段落均提及金面佛、胡一刀等。

13、《倚天屠龙记》的训练结果如下表所示：

方法一： 词向量语义距离	**赵敏**	**关联度**
	张无忌	0.4490940272808075
	周芷若	0.39407262206077576
	小昭	0.35240957140922546
	赵敏道	0.2856042683124542
	杨不悔	0.2803437411785126
	灭绝师太	0.2782444655895233
	圆真	0.24351656436920166
	鹿杖客	0.24217510223388672
	范右	0.24020634591579437
	宋青书	0.23047876358032227
方法二： 词语聚类	**词组**	**类别**
	张无忌	0
	赵敏	0
	周芷若	0
	灭绝师太	1
方法三： 段落语义关联	**所选段落关联度**	
	0.3308938443660736	

注：所选段落均提及张无忌、赵敏、倚天剑等。

14、《鸳鸯刀》的训练结果如下表所示：

方法一： 词向量语义距离	**林玉龙**	**关联度**
	任飞燕	0.7516977787017822
	二人	0.5082764029502869
	夫妻	0.46140235662460327
	骂	0.40992680191993713

	原来	0.39849066734313965
	瞎子	0.33360224962234497
	问	0.33101195096969604
	道	0.32822972536087036
	刀法	0.321713924407959
	夫妇	0.3014751970767975
方法二： 词语聚类	**词组**	**类别**
	林玉龙	1
	任飞燕	1
	袁冠南	0
	萧中慧	0
方法三： 段落语义关联	**所选段落关联度**	
	0.42547377943992615	

注：林玉龙和任飞燕是夫妻，袁冠南和萧中慧是夫妻；所选段落均提及林玉龙、任飞燕等。

15、《越女剑》的训练结果如下表所示：

方法一： 词向量语义距离	**范蠡**	**关联度**
	说道	0.7930134534835815
	说	0.7803463935852051
	阿青	0.751732587814331
	西施	0.724307119846344
	范大夫	0.7192902565002441
	姑娘	0.6781548261642456
	笑	0.6433240175247192
	一个	0.6215177774429321
	见	0.555661141872406
	大夫	0.47328436374664307
方法二： 词语聚类	**词组**	**类别**
	阿青	0

	范蠡	0
	西施	0
	勾践	1
方法三： 段落语义关联	**所选段落关联度**	
	0.849465548992157	

注：所选段落均提及范蠡、阿青、西施等。

16、《三十三剑客图》的训练结果如下表所示：

方法一： 词向量语义距离	**卢生**	**关联度**
	唐山人	0.41732776165008545
	舅舅	0.3394268751144409
	任愿	0.32607537508010864
	起来	0.3215097486972809
	身上	0.3203800320625305
	数年	0.27162671089172363
	原来	0.25599539279937744
	今日	0.24826139211654663
	哥哥	0.24468792974948883
	其实	0.2327202409505844
方法二： 词语聚类	**词组**	**类别**
	虬髯	0
	卢生	0
	唐山人	1
方法三： 段落语义关联	**所选段落关联度**	
	0.8965762853622437	

注：虬髯客是一幅图的故事，卢生和唐山人是另一幅图的故事；所选段落均提及卢生、唐山人等。

从以上实验结果可以看出，与目标词相关的角色，在计算词向量语义距离时结果都比较高；对于给定的一组词，能根据不同特征（角色或门派、主角或反派、是否伴侣等）进行比较准确的聚类；对任意选取的段落而言，相似的情节或相同

的角色越多，段落关联度越高。

Conclusion

本实验使用金庸的 16 部小说作为语料库，利用 Word2Vec 模型来进行词向量训练，并通过计算词向量之间语意距离、对某一类词语聚类、计算某些段落直接的语意关联三种方法，验证了词向量的有效性。