Report of Deep Learning for Natural Langauge Processing

基于 word2vec 的词向量训练与验证

韩子轩
18210237202@163.com

Abstract

本实验以金庸小说作为语料库，首先使用 word2vec 模型训练词向量，采用了 skip-gram 架构。之后通过计算词语的聚类与词向量之间的相似度来验证该词向量的有效性。经过实验，取得较好的验证效果。

Introduction

利用给定语料库（金庸语小说料如下链接），利用 1～2 种神经语言模型（如：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。

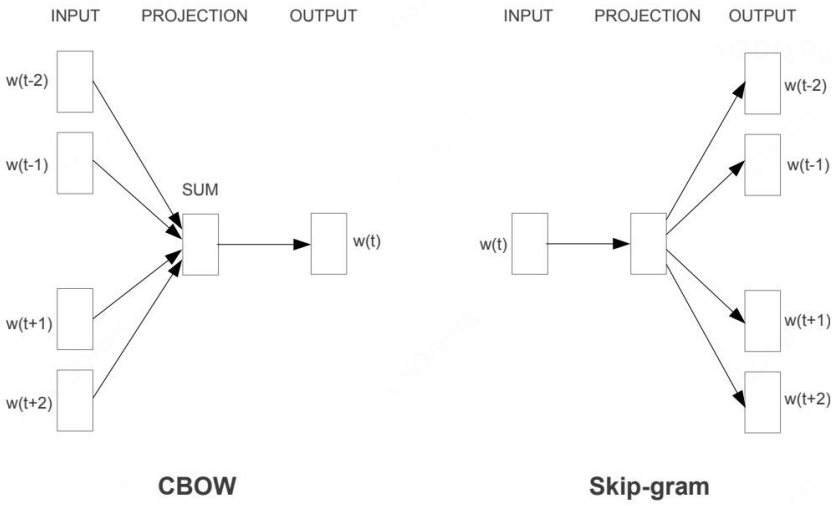
Methodology

M1: Word2vec 模型

word2vec 在 2013 年由 Google 的 Tomas Mikolov 等人提出，它主要基于分布假设，即在语料库中相似上下文中出现的单词具有相似的语义，采用浅层神经网络来生成单词的向量。word2vec 的模型可以划分为三层，输入层，隐藏层和输出层，输入层就是 one-hot 编码，输入层到隐藏层之间有一个 M*V 的矩阵(M 是词表大小，V 是向量长度)，输入层到隐藏层其实就是 one-hot 到向量的转化过程，然后隐藏层到输出层之间会有各种函数计算，也是训练的主要内容，得到词表大小的概率，训练的目标是提升正样本单词的概率，减小负样本单词的概率。

Word2vec 主要有两种主要的架构，其中本实验主要采用 skip-gram 架构：

- Continous Bag of Words(CBOW)：根据上下文来预测中心词
- Skip-gram：根据中心词预测上下文



skip-gram 是一种基于神经网络的无监督学习算法，其设计思路是根据目标词汇来预测上下文。在自然语言体系中，一个语法正确具有语义的文本中，词与词之间存在某种语言上的联系。既然句子中词与词是存在某种关联的，那么就意味着可以根据某些词来预测出与其一起出现的其他词语，被预测的词语记为 target word，target word 周围的词语记为 context word。

在给定 target words 的情况下，skip-gram 模型使得 context words 出现的概率最大，即

$$\underset{\theta}{argmax}\ p(\omega_{t-k},\omega_{t-k+1},\cdots,\omega_{t-1},\omega_{t+1},\omega_{t+2},\cdots,\omega_{t+k}|\omega_t;\theta)$$

其中θ表示权重矩阵，k表示 window size. 根据概率知识，上式可转化为：

$$\underset{\theta}{argmax}\prod_{-k\le j\le k;j\neq 0} p(\omega_{t+j}|\omega_t;\theta)$$

为计算方便，将连乘通过 log 转换成累加：

$$\underset{\theta}{argmax}\sum_{-k\le j\le k;j\neq 0} \log p(\omega_{t+j}|\omega_t;\theta)$$

在训练过程中，损失函数一般取最小值。考虑训练集中存在 T 个 target words 时，损失函数如下

$$loss=-\frac{1}{T}\sum_{t=1}^{T}\sum_{-k\le j\le k;j\neq 0} \log p(\omega_{t+j}|\omega_t;\theta)$$

在 skip-gram 中，上式中的概率通过 softmax 函数来计算，即

$$p(\omega_{t+j}|\omega_t;\theta)=\frac{\exp(\omega_{t+j}^T\omega_t)}{\sum_{i=1}^{V}\exp(\omega_i^T\omega_t)}$$

其中V代表训练数据集中 Vocabulary size。

Experimental Studies

1.数据预处理

在读取语料后，首先利用 jieba 分词对语料进行分词，去掉 txt 文本中一些无意义的广告和标点符号等内容，并将处理后的语料重新保存进新的文件夹中。

在实验过程中，注意到某些无关词语会在结果中出现，因此在进行语料处理时，去掉一些无关词语（例如，“我”、“你”、“他”、“她”、“它”等）

2.word2vec 模型训练

使用开源的 Gensim 库提供的接口来训练 Word2vec 模型，调用的函数如下：

```
model = Word2Vec(sentences=LineSentence(name), hs=1, min_count=10, window=5, vector_size=200, sg=1, epochs=200)
```

其中的参数的含义为：

sentences：可以是一个 list，对于大语料集，建议使用 BrownCorpus,Text8Corpus 或 LineSentence 构建；

hs: 如果为 1 则会采用 hierarchical softmax 技巧。如果设置为 0（defaut），则 negative sampling 会被使用；

min_count: 可以对字典做截断,词频少于 min_count 次数的单词会被丢弃掉，默认值为 5，这里的值为 10。

window：表示当前词与预测词在一个句子中的最大距离是多少，这里为 5；

vector_size：是指特征向量的维度，默认为 100；

sg： 用于设置训练算法，默认为 0，对应 CBOW 算法；sg=1 则采用 skip-gram 算法；

epochs：迭代次数，默认为 5。

3.聚类分析

模型训练完毕之后,通过 Gensim 库中给出的接口函数(例如,most_similar(),similarity()等)，输出训练之后的模型，与某个给定输入词关联度最高的词或者是给定的某两个词之间的关联性。本实验选择《天龙八部》小说为数据集，对“段誉”和“逍遥派”进行了聚类分析。结果如下：

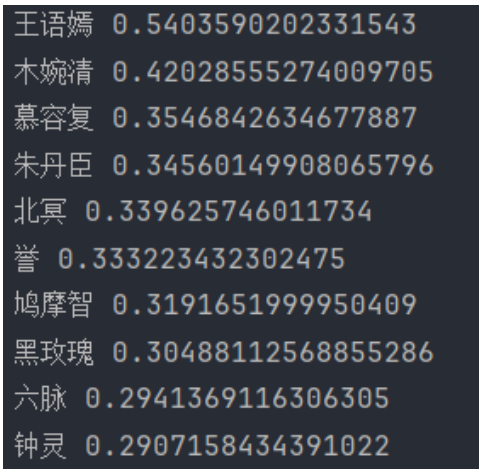


图 1 “段誉”的词语聚类和关联度

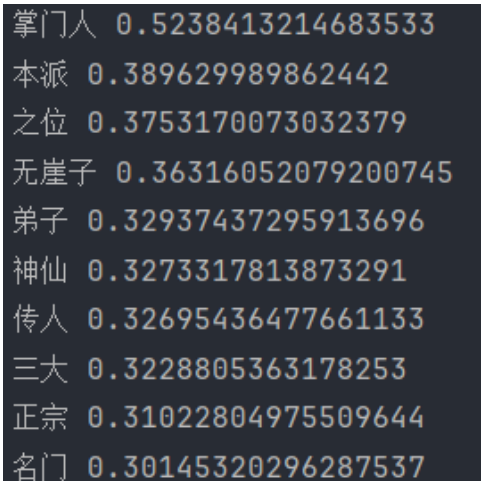


图 2“逍遥派”的词语聚类和关联度

可以发现，和“段誉”相似度最大的十个词语主要是和他关系最大的一些人物，比如王语嫣、木婉清等，同时也包含了他的武功，如北冥（神功）、六脉（神剑）。和“逍遥派”关联最大的则是掌门人、之位等，这也和小说中逍遥派围绕着掌门人位置的争夺的相关剧情相符。因此，词向量的有效性得到验证。

Conclusions

本实验使用 skip-gram 架构的 word2vec 模型训练词向量，并计算《天龙八部》中词语的聚类与词向量之间的相似度，发现人物和门派的关联度均较高，符合小说实际情况，验证了该方法的有效性。