文章编号:1005-6033(2015)02-0145-04 收稿日期:2014-11-18

Word2vec 的工作原理及应用探究

周 练

(西安电子科技大学经济与管理学院,陕西西安,710071)

摘 要:研究了 Word2vec 的工作原理及应用,明确了统计语言模型的关键问题,分析了词向量的特点,并对神经网络语言模型、Log_Linear 模型和 Log_Bilinear 模型的基本原理进行了探讨,对 Word2vec 词向量训练框架的工作原理进行了详细分析,推导出了训练模型的目标函数,介绍了 Word2vec 工程的主要文件和训练参数,并将 Word2vec 应用于中文词向量的训练。

关键词:Word2vec;词向量;统计语言模型

中图分类号:TP317 **文献标识码**:A

1 问题的提出

随着计算机应用领域的不断扩大,自然语言处理受到了人们的高度重视。机器翻译、语音识别以及信息检索等应用需求对计算机的自然语言处理能力提出了越来越高的要求。

为了使计算机能够处理自然语言,首先需要对自然语言进行建模。自然语言建模方法经历了从基于规则的方法到基于统计方法的转变[1]。从基于统计的建模方法得到的自然语言模型称为统计语言模型。有许多统计语言建模技术,包括 n-gram、神经网络以及 log_linear 模型等。在对自然语言进行建模的过程中,会出现维数灾难、词语相似性、模型泛化能力以及模型性能等问题。寻找上述问题的解决方案是推动统计语言模型不断发展的内在动力。

在对统计语言模型进行研究的背景下,Google 公司在 2013 年开放了 Word2vec[2]这一款用于训练词向量[3]的软件工具。Word2vec 可以根据给定的语料库,通过优化后的训练模型快速有效地将一个词语表达成向量形式,为自然语言处理领域的应用研究提供了新的工具。

2 相关工作

2.1 统计语言模型

统计语言模型是用于刻画一个句子出现概率的模型。给定一个由 n 个词语按顺序组成的句子 $S=(w_1,w_2,\cdots,w)$,则概率 $p(S)$ 即为统计语言模型。通过贝叶斯公式,可以将概率 $p(S)$ 进行分解,即 $p(S)=p(w)\cdot p(w_2|w)\cdot p(w_3|w_1^2)\cdots p(w_n|w_1^{n-1})$。所以,要计算一个句子出现的概率,只需要计算出在给定上下文的情况下,下一个词为某个词的概率即可,即 $p(w_i|context(w))$。当所有条件概率 $p(w_i|context(w))$ 都计算出来后,通过连乘即可计算出 $p(S)$。所以,统计语言模型的关键问题在于找到计算条件概率 $p(w_i|context(w))$ 的模型。

2.2 词向量

词向量具有良好的语义特性,是表示词语特征的常用方式。

第一作者简介:贾 茹,女,1980 年 11 月生,2007 年毕业于天津工业大学材料学专业(硕士),馆员,天津职业大学图书馆,天津市北辰区洛河道 2 号,300402.

Cleaning off the English Obstacles in Literature Retrieval Course in Higher Vocational College

JIA Ru

ABSTRACT: Higher vocational college students' English level will directly influence the teaching effects of literature retrieval course. According to the characteristics of literature retrieval course and the status of students' English level, this paper puts forward some suggestions on improving the teaching methods of literature retrieval course, which include applying the English education through in the whole course, combining the literature retrieval course with the professional English course, and carrying out the bilingual teaching if possible.

KEY WORDS: literature retrieval course; English obstacles; higher vocational college

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

根据文献[3]表述，词向量每一维的值代表一个具有一定的语义和语法上解释的特征。所以，可以将词向量的每一维称为一个词语特征。词向量具有多种形式，distributed representation 是其中一种。一个 distributed representation 是一个稠密、低维的实值向量。distributed representation 的每一维表示词语的一个潜在特征，该特征捕获了有用的句法和语义特性。可见，distributed representation 中的 distributed 一词体现了词向量这样一个特点：将词语的不同句法和语义特征分布到它的每一个维度去表示。

2.3　神经网络语言模型

2003 年，Bengio 等[4]提出一个基于 3 层神经网络的自然语言估计模型 NNLM（Neural Network Language Model），如图 1 所示。NNLM 可以计算某一个上下文的下一个词为 w_t 的概率，即 $p(w_t=i|context)$，词向量是其训练的副产物。

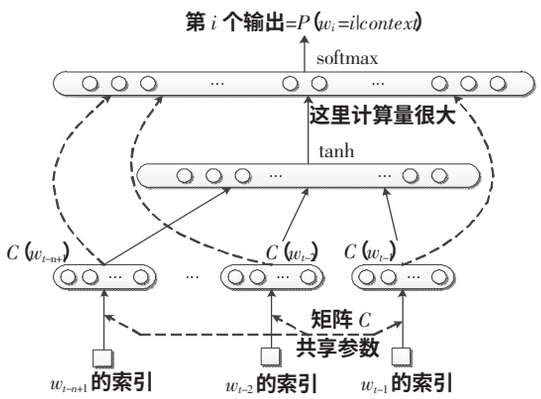

图 1　神经网络语言模型

NNLM 根据语料库 C 生成一个对应的词汇表 V。V 中的每一个词汇都对应着一个编号 i。为了确定神经网络的参数，需要通过语料库来构建训练样本并作为神经网络的输入。样本的构建过程为：对于 C 中的任意一个词 w_t，获取其上下文 $context(w)$（例如前 $n-1$ 个词），从而得到一个元组 $(context(w),w)$。以该元组作为神经网络的输入进行训练。由图 1 可知，对于每一个词，NNLM 都将其映射成一个向量 $C(w)$，该向量即为词向量。NNLM 在输出层上的 softmax 归一化函数为 $p(w_t|w_{t-1},\cdots,w_{t-n+2},w_{t-n+1})=\frac{e^{y_{w_t}}}{\sum_i e^{y_i}}$，其中 y_i 是下标为 i 的词语对应的未归一化的 log 概率值。可见，当词汇表很大时，softmax 的操作将非常耗时。为了解决这个问题，Bengio 等[4]在 NNLM 的输出层上进行层次化改进，大大提高了 NNLM 的计算效率。

层次化操作的基本思想是对词语进行分类。例如，在包含 10 000 个词的词汇表的情况下，对于每一个概率 $p(w_t|context)$，NNLM 的输出层要对 w_t 的每个可能取值进行一次归一化，每个取值总共需要计算 10 000 次。对词汇表中的词进行分类后，比如分为 100 个类别，每个类别为 100 个词。则对于 w_t 的每个可能取值，只需要进行 200 次计算即可，加速比达到 50。可以对词汇表进行更多层次的划分来提高计算的速度。

2.4　Log_Linear 和 Log_Bilinear 模型

Log_Linear 模型可用于自然语言建模。同时，Log_Linear 模型也是 Word2vec 训练模型的基础。根据文献[6]，Log_Linear 模型的组成部分包括：

⑴ 一个可能的输入集合 X；

⑵ 一个有限的标记集合 Y；

⑶ 一个指定模型中的特征和参数数目的正整数 d；

⑷ 一个将 (x,y) 映射到一个特征向量 $f(x,y)$ 的函数 $f:X\times Y\to R^d$；

⑸ 一个参数向量 $v\in R^d$。

Log_Linear 模型定义了一个条件概率，即对于任意的 $x\in X$，$y\in Y$，$p(y|x;v)=\frac{\exp(v\cdot f(x,y))}{\sum_{y'\in Y}\exp(v\cdot f(x,y'))}$。其中，$\exp(x)=e^x$，$v\cdot f(x,y)=\sum_{k=1}^{d}v_k f_k(x,y)$。

$f(x,y)$ 是一个代表 (x,y) 的特征向量，$f(x,y)$ 的每一维 $f_k(x,y)$ 代表一个特征，每一个特征与 v 的一个维度 v_k 相关联。v 中每一维的参数 v_k 需要通过训练样本训练得到。

在 Log_Linear 模型基础上，Hinton[7]提出了 Log_Bilinear 模型。Log_Bilinear 模型与 Log_Linear 模型的主要区别在于映射函数 $f(x,y)$ 部分。在 Log_Bilinear 模型中，$f(x,y)$ 将输入 (x,y) 直接映射成一个 d 维特征向量，而 Log_Bilinear 模型直接使用输入 y 对应的向量。通过借鉴神经网络语言模型中的层次化思想，Hinton 提出了层次化的 Log_Bilinear 模型，这也是 Word2vec 所使用的模型。

3　Word2vec 的工作原理

3.1　Word2vec 的训练模型

Word2vec 是 Mikolov 等[8-9]所提出模型的一个实现，可以用来快速有效地训练词向量。Word2vec 包含了两种训练模型，分别是 CBOW 和 Skip_gram，如图 2 所示。

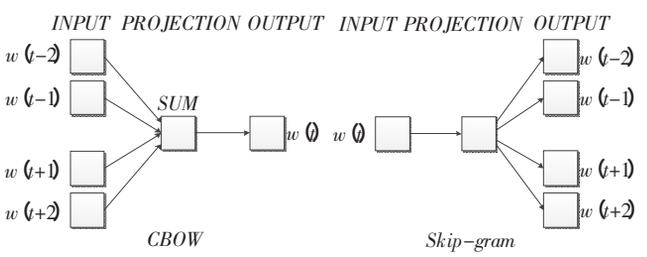

图 2　CBOW 和 Skip_gram 模型

从图 2 可以看出，CBOW 和 Skip_gram 模型均包含输入层、投影层和输出层。其中，CBOW 模型通过上下文来预测当前词，Skip_gram 模型则通过当前词来预测其上下文。Word2vec 提供了两套优化方法来提高词向量的训练效率，分别是 Hierachy Softmax 和 Negative Sampling。将训练模型和优化方法进行组合可得到 4 种训练词向量的框架，如表 1 所示。

3.2　Word2vec 训练词向量的原理

3.2.1　CBOW+HS 和 Skip_gram+HS

从 Word2vec[10] 的源码中可以归纳出 CBOW+HS 和 Skip_gram+HS 模型的训练框架，如图 3 所示。根据文献[11]，总结得到两种训练框架的共同点和区别，如表 2 所示。

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表 1　word2vec 词向量训练框架

模型	CBOW	Skip_gram
Hierachy Softmax	CBOW+HS	Skip_gram+HS
Negative Sampling	CBOW+NS	Skip_gram+NS

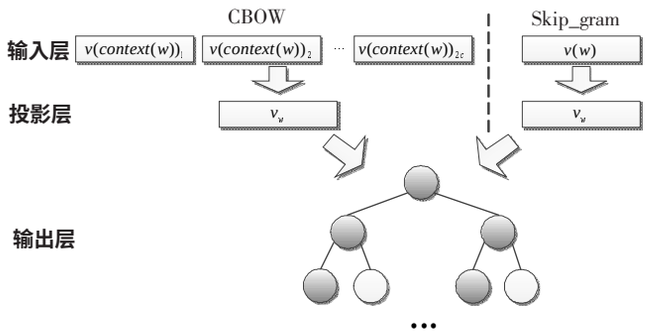

图 3　CBOW+HS 和 Skip_gram+HS 训练框架

表 2　CBOW+HS 和 Skip_gram+HS 训练框架对比

对比项	CBOW+HS	Skip_gram+HS
训练样本	已知上下文，预测下一个词（$context(w)$，w）	已知当前词，预测上下文（$context(w)$，w）
输入层	词语 w 前后各 c 个词对应的词向量	当前词 w 对应的词向量
投影层	输入层的 $2c$ 个词向量的累加	当前词 w 对应的词向量（恒等投影）
输出层	以词语在语料库中的词频作为权值构造的一棵二叉树。叶子节点对应词汇表中的所有词语。假设叶子节点为 N 个，则非叶子节点为 $N-1$ 个。叶子节点和非叶子节点均对应一个向量。其中叶子节点对应的向量即为词向量，而非叶子节点对应的向量是一个辅助向量。	
训练参数	词汇表中词语对应的词向量，输出层非叶子节点对应的辅助向量。	

在模型的训练过程中，梯度是训练参数更新的依据。为了获得梯度公式，需要构造出训练模型的目标函数。在 CBOW+HS 框架的训练中，给定一个训练样本（$context(w)$，w），则在输出层从根节点到词 w 的路径上，对于每一个非叶子节点均对应一个辅助向量 θ^w_j 和一次二分类且左右两边各对应一个哈夫曼编码 l^w_j。其中，j 表示从根节点到叶子节点 w 路径上非叶子节点的编号。对于 l^w_j，规定向左的分支对应 1，向右的分支对应 0。这样，通过若干二分类后，最终可到达叶子节点 w。可将非叶子节点向左的分支定义为负类，将向右的分支定义为正类。通过逻辑回归知识，可以得到一个二分类被分成正类的概率为 $\sigma(v^T_w\theta)=\frac{1}{1+e^{-v^T_w\theta}}$，被分成负类的概率为 $1-\sigma(v^T_w\theta)$。其中，v^T_w 是 $context(w)$ 中包含的词所对应词向量的累加和，而 θ 为非叶子节点对应的辅助向量。沿着从根节点到叶子节点 w 的路径将每个非叶子节点的二分类概率相乘即为 $p(w|context(w))$。

对于 Skip_gram+HS，给定一个训练样本（w，$context(w)$），其中 $context(w)$ 包含 $2c$ 个词。可以将通过 w 预测 $context(w)$ 的问题，即 $p(context(w)|w)$ 转换成 $2c$ 个通过 w 预测下一个词为 u 的问题 $p(u|w)$，其中 $u\in context(w)$。$p(u|w)$ 可以利用 CBOW+HS 框架的思路来解决，即将 u 视为叶子节点。不同的是，在 CBOW+HS 框架中，v^T_w 指的是 $context(w)$ 中所有词对应词向量的累加，而在 Skip_gram+HS 中指的是 w 对应的词向量。Skip_gram+HS 框架的目标函数为 $p(context(w)|w)=\prod_{u\in context(w)}p(u|w)$。

3.2.2　CBOW+NS 和 Skip_gram+NS

CBOW+HS 和 Skip_gram+HS 框架输出层的哈夫曼树构造过程相对复杂。所以，CBOW+NS 和 Skip_gram+NS 框架采用了一种替代的方法，即采用相对简单的负取样来提高词向量训练的速度。

对于 CBOW+NS 框架，已知样本（$context(w)$，w），则 w 为一个正样本，而词汇表中的其他词为负样本。可以采用不同的负采样算法来选择负样本。在确定关于 $context(w)$ 的一个非空的负样本集合 $NEG(w)$ 后，对于词汇表中的任何词 w'，若 $w'=w$，给其为 1 的标签，否则给其为 0 的标签。这样，正样本的标签为 1，负样本的标签为 0。这里的标签可类比成 CBOW+HS 框架输出层哈夫曼树中非叶子节点的左右二分类编码。优化的目标函数可表示为 $p(w)=\prod_{u\in\{w\}\cup NEG(w)}p(u|context(w))$。

对于 Skip_gram+NS 框架，已知样本（w，$context(w)$）。则对于一个上下文词 w'，当 $w'=w$，给其为 1 的标签，否则给其为 0 的标签。这里的标签同样可类比成 CBOW+HS 框架输出层哈夫曼树中非叶子节点的左右二分类编码。则优化的目标函数为 $p(w)=\prod_{w'\in context(w)}\prod_{w''\in\{w\}\cup NEG(w)}p(w''|w)$。

4　Word2vec 的应用

4.1　Word2vec 项目介绍

4.1.1　Word2vec 的工程目录

从文献 [10] 中导出的 Word2vec 项目包含了若干个 C 语言原文件和 linux 脚本。其中，与训练词向量相关的文件主要包括 Word2vec.c、demo-word.sh 和 distance.c 3 个文件。Word2vec.c 文件包含了 Word2vec 各个模型的实现，demo-word.sh 包含了模型训练需要指定的参数列表，而 distance.c 文件可以计算不同词向量间的余弦值。在 linux 平台下，通过 demo-word.sh 脚本即可启动 Word2vec 进行词向量的训练。distance.c 文件以训练好的词向量文件作为输入，可以获取与某个词语在语义上最相近的词语列表。

4.1.2　Word2vec 的训练参数

Word2vec 提供了许多超参数来调整训练过程。不同参数的选择对生成的词向量的质量以及训练的速度有影响。Word2vec 可调整的参数如表 3 所示。

4.2　利用 Word2vec 训练中文词向量

对于英文语料而言，Word2vec 可以根据词语之间的空格来识别不同的词语。但是，对于中文而言，Word2vec 不能直接识别。所以，要使用 Word2vec 来训练中文词向量，首先需要对中文语料进行分词。本文对搜狗 2012 年 6 月到 7 月的新闻数据进行分词来生成中文语料。将分好词的中文语料作为 Word2vec 的输入文件并指定合适的训练参数，即可进行中文词向量的训练，如图 4 所示。

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表 3　Word2vec 的训练参数

参数	说明	参数	说明	参数	说明
train	输入文件的路径	output	输出文件位置	size	词向量的维数
window	窗口大小	hs	是否采用 softmax 体系	negative	负样本的数量
threads	开启的线程数量	min-count	词语出现的最小阈值	alpha	学习率初始值
binary	是否用 binary 模式保存数据	cbow	是否采用 CBOW 算法	classes	输出词类别

```
Starting training using file resultbig.txt
Vocab size: 348863
Words in train file: 379380817
Alpha: 0.024113  Progress: 3.55%  Words/thread/sec: 43.21k
```

图 4　词向量的训练

通过训练得到的词向量文件，可以计算出与某个词在语义上比较相近的词语。图 5 展示了与“中国”在语义上最接近的 5 个词。

Word	Cosine distance
我国	0.540655
亚洲	0.532703
世界	0.513137
全球	0.491376
大国	0.464752

图 5　与“中国”语义上最接近的 5 个词语

5　结语

Word2vec 是一款用于训练词向量的软件工具，提供了 CBOW 和 Skip_gram 两种训练模型。结合 hierarchy softmax 和 negative sampling 优化技术，Word2vec 可以快速高效地将词语表达成向量。同时，由于词向量捕获了自然语言中词语之间的语义特征，通过保存到文件中，词向量可以供其他相关应用研究使用。Word2vec 的出现为快速获取自然语言语义特征提供了可能，从而促进了自然语言处理领域相关研究的发展。

参考文献

[1]　吴军.数学之美[M].北京：人民邮电出版社，2012.
[2]　Tomas Mikolov.Word2vec project [EB/OL]. [2014-09-18]. https://code.google.com/p/word2vec/.
[3]　Joseph Turian，Lev Ratinov，Yoshua Bengio.Word representations：a simple and general method for semi-supervised learning [G]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics.Uppsala，Sweden：[s.n]，2010：384-394.
[4]　Yoshua Bengio，Rejean Ducharme，Pascal Vincent，Christian Jauvin. A neural probabilistic language model [J].Journal of Machine Learning Research，2003(3)：1137-1155.
[5]　Morin，F，Bengio Y. Hierachical probabilistic neural network language model [C].AISTATS，Hastings，Barbados，January，06-08，2005.
[6]　Michael Collins. Log-linear models [EB/OL]. [2014-09-15]. http://www.cs.columbia.edu/~mcollins/logline ar.pdf.
[7]　Mnih A，Hinton G. Three new graphical models for statistical language modelling [G]// Proceedings of the 24th international conference on Machine learning. [S.l.]：[s.n]，2007：641-648.
[8]　Tomas Mikolov，Kai Chen，Greg Corrado，et al. Efficient estimation of word representations in vector space [EB/OL]. [2014-09-19]. http://arxiv.org/abs/1301.3781v3.
[9]　Mikolov T，Sutskever I，Chen K，et al.Distributed representations of words and phrases and their compositionality [C]//NIPS，Lake Tahoe，USA，December，05-08，2013.
[10]　Tomas Mikolov.Word2vec code [CP/OL]. [2014-09-18]. http://word2vec.googlecode.com/svn/trunk/.
[11]　皮果提.Word2vec 中的数学原理 [EB/OL]. [2014-11-15]. http://blog.csdn.net/itplus/article/details/37969519.

(责任编辑　薛培荣)

第一作者简介：周　练，男，1988 年 9 月生，现为西安电子科技大学经济管理学院情报学专业 2012 级在读硕士研究生，陕西省西安市雁塔区太白南路 2 号西安电子科技大学北校区，710071.

Exploration of the Working Principle and Application of Word2vec

ZHOU Lian

ABSTRACT：This paper studies the working principle and application of Word2vec，defines the key problems of statistical language model, analyzes the characteristics of word vector, probes into the basic principles of neural network language model，Log_Linear model and Log_Bilinear model，makes a detailed analysis on the working principle of word vector's training framework of word2vec, and derives the objective functions of the training models，and introduces the main files in Word2vec project and training parameters, and applies Word2vec into the training of Chinese word vector.

KEY WORDS：Word2vec; word vector; statistical language model

?1994-2017 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net