第 44 卷 第 2 期 计 算 机 工 程 2018 年 2 月
Vol. 44 No. 2 Computer Engineering February 2018

·人工智能及识别技术· 文章编号：1000-3428(2018)02-0233-05 文献标志码：A 中图分类号：TP391.1

主题联合词向量模型

吴旭康[1,2]，杨旭光[3]，陈园园[3]，王营冠[1]，张阅川[3]

（1. 中国科学院上海微系统与信息技术研究所，上海 200050；
2. 上海科技大学 信息科学与技术学院，上海 201210；3. 上海物联网有限公司，上海 200018）

摘 要： 当前大部分的词向量模型针对一个单词只能生成一个向量，由于单词的多义性，使用同一个向量表达不同语境下的同一个单词是不准确的。对此，提出一种新的词向量模型。使用潜狄利克雷特分布和神经网络对单词进行训练，得到单词及其主题的向量，并对两者进行线性变换得到最终的词向量。实验结果表明，该模型的准确度高于现有多向量模型。

关键词： 自然语言处理；词向量；主题模型；神经网络；哈夫曼树

中文引用格式： 吴旭康，杨旭光，陈园园，等. 主题联合词向量模型[J]. 计算机工程，2018，44(2)：233-237，270.

英文引用格式： WU Xukang，YANG Xuguang，CHEN Yuanyuan，et al. Topic Combined Word Vector Model[J]. Computer Engineering，2018，44(2)：233-237，270.

Topic Combined Word Vector Model

WU Xukang[1,2]，YANG Xuguang[3]，CHEN Yuanyuan[3]，WANG Yingguan[1]，ZHANG Yuechuan[3]

（1. Shanghai Institute of Microsystem and Information Technology，Chinese Academy of Sciences，Shanghai 200050，China；2. School of Information Science and Technology，ShanghaiTech University，Shanghai 201210，China；3. Shanghai Internet of Things，Co.，Ltd.，Shanghai 200018，China）

【Abstract】 Currently，most word vector models can build only one vector for a single word. Due to word's polysemy，it is incorrect to use one vector representing a same word under different context. This paper proposes a new word vector model. It uses latent dirichlet distribution and neural networks to train words to obtain word vectors and corresponding topic vectors. And then it applies linear transformations on them to build the final word vectors. Experimental results show that the accuracy of proposed model is high compared with current multi-vector models.

【Key words】 natural language processing；word vector；topic model；neural network；Haffman tree

DOI： 10.3969/j.issn.1000-3428.2018.02.040

0 概述

词向量[1-2]是用数学形式的向量来表达单词，可用于自然语言处理任务中的很多方面，比如命名实体识别、句子成分分析[3]、单词相似度计算等。因此，词向量的研究得到了越来越多的关注。

当前，大部分的词向量模型通常使用一个向量来表示一个单词，忽视了单词的多义性，从而削弱了词向量对单词语义表达的唯一性。为了解决这样的问题，多向量模型被提出[4-5]，该模型通过对同一个单词的不同上下文语境（context）进行聚类，针对每一个类簇生成一个词向量，然而，由于该方法过于笨拙和繁琐，于是，另一些基于神经语言模型[6-8]的词向量模型被提出，这类模型通过构造不同的神经网络，使用随机梯度下降（Stochastic Gradient Descent，SGD）或者 EM（Expectation Maximum）算法得到最终的模型参数，进而得到词向量。

然而，即使采用多向量模型，在生成多个词向量的过程中也存在一个较大缺陷——这些多向量模型认为同一个单词的不同上下文语境是独立的，不具有相关性。事实上，即使是不同的上下文语境，仍然可以有语义（Semantic）上的相似或者重叠。例如，在英语中，单词“like”在句子“I like that girl”和句子“She is my like”中，两者表达了几乎一致的信息，却完全是 2 种上下文语境。因此，把不同的语境完全的隔离开并不是完美可行的方案。于是，TWE 模型被提出，结合单词的主题信息，得到更具表达性的词向量[9]，但该模型简单地将单词和主题向量连接作为最终的词向

基金项目： 上海市自然科学基金“阵元互耦条件下基于空域稀疏的阵列测向方法研究”（15ZR1439800）；上海市科技创新行动计划项目（15DZ1100400，16511105300）。

作者简介： 吴旭康（1992—），男，硕士，主研方向为自然语言处理；杨旭光，博士；陈园园，工程师、硕士；王营冠，研究员、博士；张阅川，硕士。

收稿日期： 2016-12-30 **修回日期：** 2017-02-25 **E-mail：** wxkdesky@hotmail.com

量,在向量表达性上不够突出。紧接着,基于主题单词嵌入(Topical Word Embedding,TWE)模型改进的单词主题混合(Word-Topic Mixture,WTM)模型[10]利用潜特征狄利克雷特分布(Latent-Feature Latent Dirichlet Allocation,LFLDA)方法,计算出 TWE 生成的单词-主题向量的概率分布,并通过假设该概率分布与狄利克雷特分布(Latent Dirichlet Allocation,LDA)得到的单词-主题概率分布一致,从而得到性能更好的词向量。然而,WTM 模型需要最小化上述 2 个概率分布的 KL 散度,计算较为繁琐。

本文提出一种更加有效和灵活的多向量模型——主体联合词向量模型(Topic Combined Word Vector Model,TCV Model)。该模型的主要思想与 WTM 模型类似,利用单词所属的主题信息来表达单词的上下文语境。不同于 WTM 模型,该模型利用哈夫曼编码为每一个单词构建一个初始向量作为神经网络的输入,神经网络训练后得到每一个单词及其主题对应的向量。TCV 模型对生成的单词向量和主题向量进行归一化和线性组合,将其作为该主题下单词的词向量,并考虑将具有最大概率的一个或 2 个主题作为该单词的有效上下文,避免 WTM 模型中最小化 KL 散度的复杂计算。

1 模型描述

1.1 LDA 模型

作为非监督型主题模型,LDA 模型可以从一系列文档中,找到指定数目的主题[11]。该模型是一个词袋子(Bag of Word,BOW)模型[12-13],文本中每个词的出现都是独立的,不依赖于其他词是否出现。在 LDA 模型中,假设:1)文档蕴含多个主题,主题数量适当;2)这是一个有生成过程的概率模型,并假设每一个文档都是被生成的。

文档的生成过程为:1)随机选择一个主题分布;2)对文档里的每一个单词,首先随机从主题分布中选择一个主题,然后从相应的主题中随机选择一个单词;3)一个主题是在一个固定词库上的分布,并且,所有的主题被假定为先于文档生成。4)搜寻的主题个数需要预先指定 LDA 模型的框图如图 1 所示,假设语料库中有 M 个文档和 K 个主题。其中,α 和 β 是 Dirichlet 分布的超参数,$z_{m,n}$是文档 m 中的第 n 个单词的主题,$w_{m,n}$是第 m 个文档中第 n 个单词,M 是文档的总数目,N_m是第 m 个文档中单词总数目,$\boldsymbol{\theta}_m$和 $\boldsymbol{\Phi}_k$均为向量,$\boldsymbol{\theta}_m$表示第 m 个文档中主题 z 的概率分布$Pr(\boldsymbol{z}|\boldsymbol{m})$,向量的每一列表示每个主题在文档中出现的概率,$\boldsymbol{\Phi}_k$表示在主题 k 下单词的概率分布 $Pr(\boldsymbol{w}|z_k)$,向量的每一列表示在主题 z_k下生成每个单词的概率。

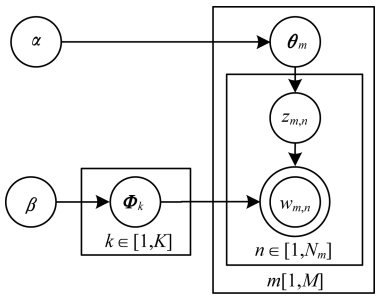

图 1 LDA 模型框图

参数为 α 的 Dirichlet 分布如下:

$$Dir(\alpha)=Pr(x|\alpha)=\frac{\Gamma\left(\sum_{i=1}^{m}\alpha_i\right)}{\prod_{i=1}^{m}\Gamma(\alpha_i)}\prod_{i=1}^{m}x_i^{a_i-1} \tag{1}$$

$$\sum_{i=1}^{m}x_i=1,x_i\geqslant 0 \tag{2}$$

其中,$\Gamma(\ )$是 Gamma 分布,x_i表示词库中被观察到的单词 i 的概率。Dirichlet 分布是多项分布的共轭先验分布。

LDA 模型的具体实施过程如下:通过对参数为 α 的 Dirichlet 分布取样生成文档 m 的主题分布 $\boldsymbol{\theta}_m$,接着从主题的多项分布 $\boldsymbol{\theta}_m$取样生成文档 m 的第 n 个单词的主题 $z_{m,n}$,并从参数为 β 的 Dirichlet 分布取样生成主题 $z_{m,n}$的单词分布 $\boldsymbol{\Phi}z_{m,n}$,最后从单词的多项分布 $\boldsymbol{\Phi}z_{m,n}$中取样最终生成的单词 $\boldsymbol{w}_{m,n}$。如此循环直到所有的文档都被生成。

通过 LDA 模型计算,每一个单词将会被赋予具有一定概率的主题标签,同时每一个主题都能通过概率排序找到最接近该主题含义的单词。如表 1 所示,在 4 个主题下各自最接近的 5 个单词(主题的名字是在观测完对应的单词分布后人为添加的)。

表 1 LDA 模型不同主题下的单词

艺术	支出	孩子	教育
New	Million	Children	School
Film	Tax	Women	Student
Show	Program	People	Schools
Music	Budget	Child	Education
Movie	Billion	Years	Teachers

1.2 Skip-Gram 模型

在谷歌正式推出词向量工具 Word2Vec 之后,Skip-Gram 模型成为一个用于生成词向量的热门简化版神经语言模型[11]。Skip-Gram 模型的目标在于预测给定单词的上下文单词。其神经网络结构如图 2所示。

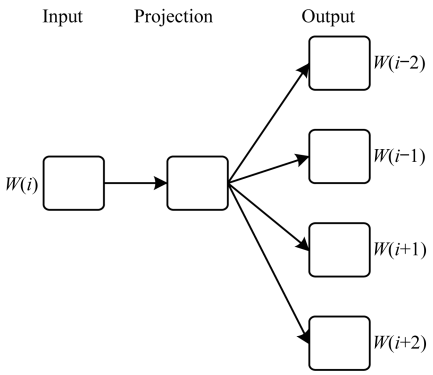

图 2　Skip-Gram 模型

神经网络的输入是一个通过哈夫曼树构造的初始词向量。哈夫曼树的节点权值由词频(Word frequency)决定。给定一个单词序列 $W=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_N\}$,该模型的目标函数是最大化平均对数概率,如下:

$$L=\frac{1}{N}\sum_{i=1}^{N}\sum_{|c|\leqslant k,c\neq 0}\mathrm{lb}\ Pr(\boldsymbol{w}_{i+c}\mid\boldsymbol{w}_i) \tag{3}$$

其中,c 表示范围为 k 的单词的上下文,N 表示训练单词的个数。为了计算 $Pr(\boldsymbol{w}_{i+c}\mid\boldsymbol{w}_i)$,引入了 Softmax 函数:

$$Pr(\boldsymbol{w}_O\mid\boldsymbol{w}_I)=\frac{\exp(\boldsymbol{v}'^{\mathrm{T}}_{wO}\boldsymbol{v}_{wI})}{\sum_{w=1}^{W}\exp(\boldsymbol{v}'^{\mathrm{T}}_{w}\boldsymbol{v}_{wI})} \tag{4}$$

其中,$\boldsymbol{v}_w,\boldsymbol{v}'_w$分别是输出、输入的词向量。$W$ 是词库的大小。由于一般的 Softmax 函数在采用随机梯度下降算法时,W 很大导致计算复杂度极高,因此层次 Softmax 将会被应用于 Skip-Gram 模型[11]。

层次 Softmax 模型主要的优势在于将计算的节点数从原来的 W 个下降至 lb W 个。该模型使用二叉树的方式呈现输出层,即将 W 个单词作为 W 颗树的叶子节点,每棵树的任一节点的分支代表一次二分类过程。这个处理过程,实际上是通过随机行走的方式为每一个单词赋予一个概率。

详细来说,每一个单词 wd 都可以找到一条从根节点出发的路径。令 $nd(wd,j)$ 表示从根节点到单词 wd 的路径上的第 j 个节点,令 $L(wd)$ 代表该路径的长度,即 $nd(wd,L(wd))=wd$。更进一步,令 $child(nd)$ 代表节点 nd 的任一固定子节点,运算符[ ]的含义是,如果 x 为真,那么[x]为 1,x 为否,则[x]为 −1。那么,层次 Softmax 函数可以表达为:

$$Pr(\boldsymbol{w}_O\mid\boldsymbol{w}_I)=\prod_{j=1}^{L(w)-1}\sigma(chk(wd,j)\cdot\boldsymbol{v}'^{\mathrm{T}}_{nd(wd,j)}\boldsymbol{v}_{wdI}) \tag{5}$$

$$chk(wd,j)=[nd(wd,j+1)=child(nd(wd,j))] \tag{6}$$

其中,$\sigma(x)=1/(1+\exp(-x))$。层次化 Softmax 在二叉树中为单词 wd 生成一个词向量 $\boldsymbol{v}_{wd}$,同时为每一个内部节点 nd 生成一个向量 $\boldsymbol{v}'_{nd}$。

1.3　主题联合词向量模型

主题联合词向量模型依赖于每个单词的主题信息。因此,首先通过 LDA 模型获取主题并对单词 $\boldsymbol{w}_i$ 标记一定数量的主题 $z_i\in T_s$。在主题标记完成后,每个单词将会有 2 个 id,分别是单词 id 和主题 id。接着,对语料库中的每一个单词构建哈夫曼树,并进行编码,作为 Skip-Gram 模型的初始单词输入,同时,为其对应的主题 id 随机生成主题编码,作为 Skip-Gram 模型的初始主题输入。主题联合词向量模型如图 3 所示。对一个单词 $\boldsymbol{w}_i$ 及其对应的主题 $\boldsymbol{z}=\{z_1,z_2,\cdots,z_j,\cdots,z_N\}$,该模型的目标函数是最大化平均对数概率:

$$L=\frac{1}{N}\sum_{i=1}^{N}\sum_{|c|\leqslant k,c\neq 0}\mathrm{lb}(Pr(\boldsymbol{w}_{i+c}\mid\boldsymbol{w}_i)Pr(\boldsymbol{w}_{i+c}\mid z_i)) \tag{7}$$

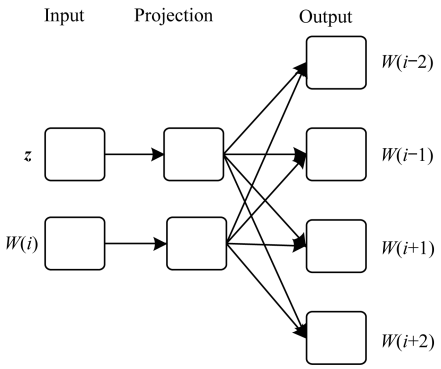

图 3　TCV 模型

当该模型训练完成后就得到单词向量 $\boldsymbol{w}_i$和主题向量 z_i。为计算的简便,本模型要求单词向量和主题向量拥有同样的维度,比如,$\boldsymbol{w}_i=\boldsymbol{W}_{1\times 200}$,$z_i=\boldsymbol{Z}_{1\times 200}$。

为了将主题信息应用于单词,从而构建更具表达性的词向量,该模型首先对主题向量进行归一化:

$$Normal(z_i)=\frac{z_i}{\max Abs(z)} \tag{8}$$

然后,对两者施加一个线性变换,得到最后的词向量 $\boldsymbol{W}_z$:

$$\boldsymbol{W}_z=Norm(z_i)(Normal(\boldsymbol{w}_i)\oplus Normal(z_i)) \tag{9}$$

其中,$Norm(z_i)$是主题向量 z_i的二范数,⊕表示将 2 个长度分别为 x,y 的向量合并成长度为 $x+y$ 的向量。这里对主题向量进行归一化的原因是相比于单词向量,主题向量各个元素的值过小(主题向量在 0.01 ~ 0.2 的范围,单词向量在 0.1 ~ 0.9 的范围)。同时,对词向量做归一化,并将两者连接成一个向量,然后用各个元素乘以主题向量的模,这个操作可以在一定程度上让主题向量和单词向量趋向统一,更好地结合两者的信息。

根据最终得到的词向量 $\boldsymbol{W}_z$,主题联合词向量模型即可用于单词相似度测试,不同于传统的单词相似度检测,该测试需要基于单词上下文。给定一个单词 w_i及其上下文 c_i,LDA 模型可以根据[8,14]进行主题概率分布的推断,即 $Pr(z|w_i,c_i)=Pr(z|c_i)Pr(w_i|z)$。因此,每一个单词将会被标记多个主题。一个很直接的上下文词向量如下:

$$\boldsymbol{W}_{c_i}^z=\sum_{z\in T_s}Pr(z|w_i,c_i)\boldsymbol{W}_z \tag{10}$$

其含义是在模型得到的所有可能的主题中,将每个主题的概率作为权重,进行加权求和,得到最终的上下文词向量。然而,对于一个给定上下文语境的单词,其语义信息大致是明确的,不会包含太多主题。因此,主题联合词向量模型接受 t 个具有最大概率的主题作为单词的主题候选,其余主题作为噪声处理。那么,新的上下文词向量表达为:

$$\boldsymbol{W}_c^z=\sum_{z\in\{1,2,\cdots,t\}}Pr(z|w_i,c)\boldsymbol{W}_z+\sigma\boldsymbol{W}_z \tag{11}$$

其中,$\sigma\boldsymbol{W}_z$被定义为噪声项,σ 为噪声系数,为了计算的简便,噪声系数在这里被设定为 0。考虑一个具有 10 个主题的单词,它唯一可能出现的场景就是字典。在一个现实的语境中,人们总会将该单词限定在某几种语义之下。10 种主题的情况是不可能出现的。因此,本文设定 t 的最大值为 2。

那么,给定一对单词及其对应的上下文(w_i,c_i)、(w_j,c_j),本模型采用余弦相似度[15]计算其词向量相似度,如下:

$$Sim(w_i,c_i,w_j,c_j)=Sim(\boldsymbol{W}_{c_i}^{z_i},\boldsymbol{W}_{c_j}^{z_j})=\frac{\boldsymbol{W}_{c_i}^{z_i}\cdot\boldsymbol{W}_{c_j}^{z_j}}{|\boldsymbol{W}_{c_i}^{z_i}||\boldsymbol{W}_{c_j}^{z_j}|} \tag{12}$$

结合式(11)和式(12),根据 AVGSimC 公式[4]计算上下文词向量的相似度 S,如下:

$$S=\sum_{z\in T_s}\sum_{z'\in T_{s'}}Pr(z|w_i,c_i)Pr(z'|w_j,c_j)Sim(\boldsymbol{W}_z,\boldsymbol{W}_z') \tag{13}$$

其中,$\boldsymbol{W}_z$为单词 w_i对应的词向量,$\boldsymbol{W}_z'$为单词 w_j对应的词向量,$T_s=\{z_1,z_2,\cdots,z_i\}$,$T_s'=\{z_1',z_2',\cdots,z_i'\}$,$i\leqslant t$。

2 实验结果

在本节中,分别从上下文单词相似度、文本分类 2 个自然语言处理任务进行实验,对比评估多种相关模型的性能。

2.1 数据集

2.1.1 单词相似度实验

传统的单词相似度实验,通常选用 WordSim353、MC、RG 等数据集作为实验对象,但是,这些数据集都忽略了单词的上下文,不适合本文的单词相似度实验。因此,在本文中采用上下文单词相似度检测数据集 SCWS,该数据集中包含 2003 对单词,每一对单词都有各自的上下文,同时,每一对单词都有 10 个人为标记的对两个单词相似度的打分,打分原则是依靠人对每一个单词在该上下文下的语义的理解打出一个 0~10 之间的分数。本文对这 10 个打分取截尾平均数作为最终的参考打分。由于余弦相似度可以为负值,这与人为的打分(均为正值)不一致,因此本文将使用以下公式,将模型得到的打分结果x_i转化为正值。

$$Score_i=\frac{x_i-min}{max-min} \tag{14}$$

其中,min 和 max 是模型输出的所有相似度结果的最大值和最小值。并且,考虑到模型计算得到的相似度数值与人为打分数值差异较大,本文采用斯皮尔曼相关系数作为最终模型打分与人为打分的接近程度。

本文使用维基百科在 2010 年 4 月的数据作为训练库。由于训练库数据巨大(训练数据有 465 万个文档),需要进行一些必要的预处理过程,例如格式化,停用词(stop word),词干提取(stemming)。该实验中将 LDA 模型的主题数目设置为 200,迭代次数设置为 100。当开始训练主题联合词向量模型时,默认窗口大小设置为 5,单词向量与主题向量维度均设置为 200。整个实验过程在一台 Inteli7 处理器、64 GB内存的工作站上进行,操作系统为 Ubuntu 14.04,训练时间为 15 d。

本文将 TCV 模型与 C&W 模型、TFIDF/Pruned TFIDF(S/M)模型[4]、经典多向量模型(例如 Huang 模型等)[6,8]、LDA 模型、Skip-Gram 模型进行比较。其中,C&W 模型和词频逆文档频(Term Frequency Inverse Document Frequency,TFIDF)模型都是单向量模型,前者不考虑任何的上下文信息,后者将单词前后的 10 个单词作为上下文并以 TFIDF 作为权重因子。Pruned TFIDF 模型也是一个单向量模型,它通过精简 TFIDF 的单词数量,将上下文中具有较低 TDIDF 值的单词去除从而提高了性能。经典多向量模型通过对不同上下文的聚类或者结合一个神经语言模型来得到最终的向量。

2.1.2 文本分类实验

本文进行文本分类实验采用的数据集是 20NewsGroups。该数据集中有数千个标记了主题的文档。本文使用其中 60% 的数据作为训练集,剩余数据作为测试集。为了能够提取文档的特征,本文引入文档向量:

$$\boldsymbol{D}=\sum_{w\in d}Pr(w|d)W_c^z \tag{15}$$

其中,d 是一个拥有一定数量主题的文档,w 是文档中的单词。为简化单词 w 出现在文档 d 中的概率 $Pr(w|d)$,使用简单高效的单词的 TFIDF 权重作为 $Pr(w|d)$。文档特征即为所有词向量的加权求和。当文档特征提取之后,使用线性支持向量分类器(Support Vector Classifier,SVC)得到最终的分类结果。在这个实验中,将 TCV 模型与 BOW 模型、LDA 模型、Skip-Gram 模型和 WTM 模型进行比较。在 BOW 模型中,依然使用 TFIDF 作为权重。LDA 模型则使用推断的主题分布来表示文档。在Skip-Gram模型中,为每个单词生成向量之后,将所有的单词向量按元素求平均,以此作为该文档的文档向量。WTM 模型的所有参数参照文献[10]说明进行设置。

2.2　结果分析

单词相似度实验结果如表 2、表 3 所示,其中,斯皮尔曼相关系数 ρ 越大,代表模型对相似度计算的结果越好。相似度比较结果被分成了 2 组——单向量组和多向量组。从实验结果中可以看到,在相似度检测实验中,主题联合的词向量模型的结果优于所有的单向量模型和多向量模型,特别是当 $t=1$ 的时候,达到了 66.9% 。

表 2　SCWS 数据集上单模型向量斯皮尔曼相关系数　%

模型-单向量	ρ
词频逆文档频率模型	26.3
潜狄利克雷特分布模型	56.9
C&W 模型	57.0
精简 TFIDF-S 模型	62.5
Skip-Gram 模型	65.7

表 3　SCWS 数据集上多模型向量斯皮尔曼相关系数　%

模型-单向量	ρ
精简 TFIDF-M 模型	60.5
Huang 模型	65.3
TCV($t=2$)模型	65.1
TCV($t=1$)模型	66.9

文本分类实验结果如表 4 所示,结果显示 TCV 模型在精度、召回率、F1 综合指标中都优越于传统模型和 WTM 模型。

表 4　文本分类实验结果　%

模型	精度	召回率	F1 综合指标
词袋子模型	79.5	79.0	79.0
狄利克雷特分布模型	70.6	70.5	70.2
Skip-Gram 模型	75.0	74.3	74.2
WTM 模型	80.9	79.5	79.2
TCV 模型	83.0	82.0	82.0

本文所提出的主题联合词向量模型,在单词向量的基础上融入具有最大概率的主题信息,可以更好地表达单词。相比于其他模型,本文模型有 3 个显著的优点:1)传统的多向量模型,对一个单词,只能生成有限数量的词向量,然而本模型却可以提取数百个主题,针对不同的主题,生成数百个词向量,在单词的呈现上更加灵活。2)传统的多向量模型通过对上下文聚类来生成不同的向量,却忽视不同上下文之间的语义交叠,而本文所提出的模型,通过依赖主题信息来生成向量,各个主题信息之间,本身就有一定程度的语义交叠,因此,本模型可以弥补将不同的上下文完全隔离的缺陷。3)WTM 模型需要最小化 KL 散度,计算较为繁琐,而本模型通过选取概率最大的一个或 2 个主题作为单词的主题,简化了计算。

3　结束语

本文提出一种新的词向量生成模型——主题联合词向量模型,能够为特定上下文语境下的单词表达以主题为特征的语义。相比传统的对单词上下文进行聚类的多向量模型,主题联合词向量模型可以结合大量的主题信息来构建基于上下文的词向量,使得生成的词向量蕴含特定主题。实验结果表明,该模型在多语境场景中具有较好的鲁棒性。由于综合了多个其他模型,后期需要对该模型进行精简,以提升整体运行速度。

参考文献

[1] TURIAN J, RATINOV L, BENGIO Y. Word Representations: A Simple and General Method for Semi-supervised Learning[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden: [s. n.], 2010: 384-394.

[2] 冯　冲,石　戈,郭宇航,等. 基于词向量语义分类的微博实体链接方法[J]. 自动化学报,2016,42(6):915-922.

[3] WANG Yiou, JUN'ICHI K Y T, TSURUOKA Y, et al. Improving Chinese Word Segmentation and POS Tagging with Semi-supervised Methods Using Large Auto-analyzed Data[C]//Proceedings of IJCNLP'11. New York, USA: [s. n.], 2011: 309-317.

[4] 李　华,屈　丹,张文林,等. 结合全局词向量特征的循环神经网络语言模型[J]. 信号处理,2016,32(6):715-723.

[5] REISINGER J, MOONEY R J. Multi-prototype Vector-space Models of Word Meaning[C]//Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. New York, USA: ACM Press, 2010: 109-117.

(下转第 270 页)

时跟踪。但由于本文算法使用改进的 SIFT 特征来解决跟踪窗口更新问题,使算法处理速度有所下降,因此下一步将重点研究如何加快处理速度。

参考文献

[1] 闫庆森,李临生,徐晓峰,等. 视频跟踪算法研究综述[J]. 计算机科学,2013,40(6A):204-209.

[2] COMANICIU D, RAMESH V, MEER P. Kernel-based Objects Tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2003,25(5):564-577.

[3] BRADSKI G R. Real Time Face and Object Tracking as a Component of a Perceptual User interface [C]// Proceedings of the 4th IEEE Workshop on Applications of Computer Vision. Washington D. C., USA: IEEE Press, 1998:214-219.

[4] HUANG S L, HONG J X. Moving Object Tracking System Based on Camshift and Kalman Filter[C]// Proceedings of the 2011 International Conference on Consumer Electronics, Communications and Networks. Washington D. C., USA: IEEE Press, 2011:1423-1426.

[5] VADAKKEPAT P, LIU J. Improved Particle Filter in Eensor Fusion for Tracking Randomly Moving Object[J]. IEEE Transactions on Instrumentation and Measurement, 2006,55(5):1823-1832.

[6] AVIDAN S. Support Vector Tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004,26(8):1064-1072.

[7] ZHANG K H, ZHANG L, YANG M H. Real-time Compressive Tracking[C]//Proceedings of 2012 European Conference on Computer Vision. New York, USA: ACM Press,2012:864-877.

[8] 王 权,刘侍刚,彭亚丽,等. 基于 SIFT 的压缩跟踪算法[J]. 现代电子技术,2014(15):62-64.

[9] 钟 权,周 进,崔雄文. 融合 SIFT 特征的压缩跟踪算法[J]. 光电工程,2015,42(2):66-72.

[10] 朱周元,张 超,吴小培,等. 尺度自适应的压缩跟踪算法[J]. 计算机工程与应用,2013,52(14):180-185.

[11] 李庆斌,朱国庆,周 妍,等. 基于特征在线选择的目标压缩跟踪算法[J]. 自动化学报,2015,41(11): 1961-1970.

[12] 景 静,徐光柱,雷帮军,等. 一种基于压缩域的实时跟踪改进算法[J]. 计算机工程,2014,40(4):170-174.

[13] ACHLIOPTAS D. Database-friendly Random Projections: Johnson-linden Stratus with Binary Coins[J]. Journal of Computer and System Sciences,2003,66(4):671-687.

[14] DIACONIS P, FREEDMAN D. Asymptotics of Graphical Projection Pursuit[J]. Annals of Statistics,1984,12(3): 793-815.

[15] COMANICIU D, RAMESH V, MEER P. Real-time Tracking of Non-rigid Objects Using Mean Shift[C]// Proceedings of 2000 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press,2000:142-149.

[16] 耿 磊,王学斌,肖志涛,等. 结合特征筛选与二次定位的快速压缩跟踪算法[J]. 自动化学报,2016, 42(9):1421-1431.

[17] 王永明,王贵锦. 图像局部不变性特征与描述[M]. 北京:国防工业出版社,2010.

[18] 曾 峦,顾大龙. 一种基于扇形区域分割的 SIFT 特征描述符[J]. 自动化学报,2012,38(9):1513-1519.

编辑 吴云芳

(上接第 237 页)

[6] HUANG E H, SOCHER R, MANNING C D, et al. Improving Word Representations via Global Context and Multiple Word Prototypes[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. New York, USA: ACM Press,2012:873-882.

[7] BENGIO Y, DUCHARME R, VINCENT P, et al. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research,2003,3:1137-1155.

[8] TIAN Fei, DAI Hanjun, BIAN Jiang, et al. A Probabilistic Model for Learning Multi-prototype Word Embeddings [C]//Proceedings of COLING' 14. New York, USA: [s. n.],2014:151-160.

[9] LIU Yang, LIU Zhiyuan, CHUA T S, et al. Topical Word Embeddings[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, USA: [s. n.],2015:2418-2424.

[10] FU Xianghua, WANG Ting, LI Jing, et al. Improving Distributed Word Representation and Topic Model by Word-topic Mixture Model[C]//Proceedings of the 8th Asian Conference on Machine Learning. Hamilton, New Zealand: [s. n.],2016:190-205.

[11] MIKOLOV T, SUTSKEVER I, CHEN Kai, et al. Distributed Representations of Words and Phrases and Their Compositionality[C]//Proceedings of Advances in Neural Information Processing Systems. New York, USA: [s. n.],2013:3111-3119.

[12] GUTHRIE D, ALLISON B, LIU Wei, et al. A Closer Look at Skip-Gram Modelling[C]//Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, Italy: [s. n.],2006:1222-1225.

[13] WALLACH H M. Topic Modeling: Beyond Bag-of-words [C]//Proceedings of the 23rd International Conference on Machine Learning, New York, USA: ACM Press,2006:977-984.

[14] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003,3:993-1022.

[15] TATA S, PATEL J M. Estimating the Selectivity of TF-IDF Based Cosine Similarity Predicates [J]. ACM Sigmod Record,2007,36(2):7-12.

编辑 刘 冰