DOI:10.11992/tis.201706012
网络出版地址:http://kns.cnki.net/kcms/detail/23.1538.TP.20170831.1051.002.html

融合语义信息的矩阵分解词向量学习模型

陈培,景丽萍
(北京交通大学 交通数据分析与挖掘北京市重点实验室,北京 100044)

摘 要:词向量在自然语言处理中起着重要的作用,近年来受到越来越多研究者的关注。然而,传统词向量学习方法往往依赖于大量未经标注的文本语料库,却忽略了单词的语义信息如单词间的语义关系。为了充分利用已有领域知识库(包含丰富的词语义信息),文中提出一种融合语义信息的词向量学习方法(KbEMF),该方法在矩阵分解学习词向量的模型上加入领域知识约束项,使得拥有强语义关系的词对获得的词向量相对近似。在实际数据上进行的单词类比推理任务和单词相似度量任务结果表明,KbEMF 比已有模型具有明显的性能提升。

关键词:自然语言处理;词向量;矩阵分解;语义信息;知识库
中图分类号:TP391 **文献标志码**:A **文章编号**:1673-4785(2017)05-0661-07

中文引用格式:陈培,景丽萍.融合语义信息的矩阵分解词向量学习模型[J]. 智能系统学报, 2017, 12(5): 661-667.
英文引用格式:CHEN Pei, JING Liping. Word representation learning model using matrix factorization to incorporate semantic information [J]. CAAI transactions on intelligent systems, 2017, 12(5): 661-667.

Word representation learning model using matrix factorization to incorporate semantic information

CHEN Pei, JING Liping
(Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China)

Abstract: Word representation plays an important role in natural language processing and has attracted a great deal of attention from many researchers due to its simplicity and effectiveness. However, traditional methods for learning word representations generally rely on a large amount of unlabeled training data, while neglecting the semantic information of words, such as the semantic relationship between words. To sufficiently utilize knowledge bases that contain rich semantic word information in existing fields, in this paper, we propose a word representation learning method that incorporates semantic information (KbEMF). In this method, we use matrix factorization to incorporate field knowledge constraint items into a learning word representation model, which identifies words with strong semantic relationships as being relatively approximate to the obtained word representations. The results of word analogy reasoning tasks and word similarity measurement tasks obtained using actual data show the performance of KbEMF to be superior to that of existing models.

Keywords: natural language processing; word representation; matrix factorization; semantic information; knowledge base

词向量是单词在实数空间所表示的一个低维连续向量,它能够同时捕获单词的语义信息和语法信息。近年来,词向量已被广泛地应用于各种各样的自然语言处理任务中[1-5],如命名实体识别、情感分析、机器翻译等。在处理上述任务的过程中通常需要用到更大单位级别(如短语、句子、段落、篇章)的向量表示,这些向量则可以由词向量组合获得。因此学习优质的词向量非常重要。

现有的词向量学习方法是利用单词的上下文信息预测该单词含义,并且使上下文信息相似的单词含义也相似,因此对应的词向量在空间距离上更

收稿日期:2017-06-06. 网络出版日期:2017-08-31.
基金项目:国家自然科学基金项目(61370129,61375062,61632004).
通信作者:景丽萍.E-mial: lpjing@bjtu.edu.cn.

靠近。现有的词向量学习方法大致可以分为基于神经网络学习词向量和基于矩阵分解学习词向量。基于神经网络学习词向量是根据上下文与目标单词之间的关系建立语言模型,通过训练语言模型获得词向量[6-12]。但有效词向量的获取是建立在训练大规模文本语料库的基础上,这无疑使计算成本很高。近年来提出的 CBOW 和 skip-gram 模型[11]去除了神经网络结构中非线性隐层,使算法复杂度大大降低,并且也获得了高效的词向量。CBOW 根据上下文预测目标单词,skip-gram 根据目标单词预测上下文单词。基于矩阵分解的词向量学习模型[13-15]是通过分解从文本语料库中提取的矩阵(如共现矩阵或由共现矩阵生成的 PMI 矩阵)得到低维连续的词向量,并且文献[13]和文献[14]证明了矩阵分解的词向量学习模型与 skip-gram 完全等价。

上述模型学习的词向量已被有效地应用于自然语言处理任务中,然而这些模型在学习词向量的过程中仅使用了文本语料库信息,却忽略了单词间的语义信息。一旦遇到下列情形很难保证所得词向量的质量:1)含义不同甚至完全相反的单词(good/bad)往往出现在相似的上下文中,那么它们的词向量必然十分相似,这明显与现实世界相悖;2)对于两个含义相似的单词,其中一个出现在语料库中的次数极少,另外一个却频繁出现,或者它们出现在不同的上下文中,那么最终它们学得的词向量会有很大差别;3)大量上下文噪音的存在使学得的词向量不能准确反映出单词间的真实关系,甚至会误导整个词向量的训练过程。

为解决上述问题,本文考虑从领域知识库提取语义信息并融入到词向量学习的过程中。这会给词向量的学习带来下列优势。

首先,知识库明确定义了单词的语义关系(knife/fork 都属于餐具,animal/dog 具有范畴包含关系等),引入这些语义关系约束词向量的学习,使学到的词向量具有更准确的关系。另外,相似单词出现在不同的上下文中或者出现频次存在较大差异带来的词向量偏差问题,都可以通过知识库丰富的语义信息予以修正。再者,知识库是各领域的权威专家构建的,具有更高的可靠性。因此,引入语义信息约束词向量的学习是很有必要的。

目前融合语义信息学习词向量已有一些研究成果。Bian 等[16]利用单词结构信息、语法信息及语义信息学习词向量,并取得了良好的效果。Xu 等[17]分别给取自于知识库的两类知识信息(R-NET 和 C-NET)建立正则约束函数,并将它们与 skip-gram 模型联合学习词向量,提出了 RC-NET 模型。Yu 等[18]将单词间的语义相似信息融入到 CBOW 的学习过程中,提出了高质量的词向量联合学习模型 RCM。Liu 等[19]通过在训练 skip-gram 模型过程中加入单词相似性排序信息约束词向量学习,提出了 SWE 模型,该模型通过单词间的 3 种语义关系,即近反义关系、上下位关系及类别关系获取单词相似性排序信息。Faruqui 等[20]采用后处理的方式调整已经预先训练好的词向量,提出了 Retro 模型,该模型可以利用任意知识库信息调整由任意词向量模型训练好的词向量,而无需重新训练词向量。

以上研究都是通过拓展神经网络词向量学习模型构建的。与之不同,本文提出的 KbEMF 模型是基于矩阵分解学习词向量。该模型以 Li 等[13]提出的 EMF 模型为框架加入领域知识约束项,使具有较强语义关系的词对学习到的词向量在实数空间中的距离更近,也就是更加近似。与 Faruqui 等采用后处理方式调整训练好的词向量方式不同,KbEMF 是一个同时利用语料库和知识库学习词向量的联合模型,并且在单词类比推理和单词相似度量两个实验任务中展示了它的优越性。

1 矩阵分解词向量学习模型相关背景

KbEMF 模型是通过扩展矩阵分解词向量学习模型构建的,本节介绍有关矩阵分解学习词向量涉及的背景知识。

矩阵分解 给定一个矩阵 $\boldsymbol{X}$,矩阵分解的目标在于找到两个低秩的矩阵 $\boldsymbol{Y}$ 和 $\boldsymbol{Z}$,使得 $\boldsymbol{X}\approx\boldsymbol{YZ}$,因此矩阵分解的目标函数可以用 $\min\limits_{\boldsymbol{Y},\boldsymbol{Z}}\zeta(\boldsymbol{X},\boldsymbol{YZ})$ 公式来表示,其中 ζ 表示将矩阵 $\boldsymbol{X}$ 近似分解为 $\boldsymbol{Z}$ 与 $\boldsymbol{Y}$ 的乘积造成的损失。不同的损失函数得到不同的矩阵分解模型,例如,非负矩阵分解[21]、概率矩阵分解[22]、最大边界矩阵分解[23]。

共现矩阵 对于一个特定的训练语料库 $\boldsymbol{T}$,$\boldsymbol{V}$ 是从该语料库中提取的全部单词生成的词汇表,当上下文窗口设定为 L 时,对任意的 $w_i\in\boldsymbol{V}$,它的上下文单词为 $w_{i-L},\cdots,w_{i-1},w_{i+1},\cdots,w_{i+L}$,则共现矩阵 $\boldsymbol{X}$ 的每个元素值 $\#(w,c)$ 表示 w 和 c 的共现次数,即上下文单词 c 出现在目标单词 w 上下文中的次数,$\#(w)=\sum_{c\in V}\#(w,c)$ 表示出现在 w 上下文中全部 c 的次数。同样地,$\#(c)=\sum_{w\in V}\#(w,c)$ 表示 c 作为上下文出现在语料库中的次数。

EMF 模型 skip-gram 模型学得的词向量在多项自然语言处理任务中都取得了良好的表现,却没有清晰的理论原理解释。由此,EMF 从表示学习的角度出发,重新定义了 skip-gram 模型的目标函数,将其精确地解释为矩阵分解模型,把词向量解释为 softmax 损失下显示词向量 d_w 关于表示字典 $\boldsymbol{C}$ 的一

个隐表示,并直接显式地证明了skip-gram就是分解词共现矩阵学习词向量的模型。这一证明为进一步推广及拓展 skip-gram 提供了坚实理论基础。EMF 目标函数用(1)式表示:

$$\min_{W,C}\zeta(X,WC) = -\mathrm{tr}(X^{\mathrm{T}}C^{\mathrm{T}}W) + \sum_{w\in V}\ln\Big(\sum_{X'_w\in S_w}\mathrm{e}^{X'^{\mathrm{T}}_w C^{\mathrm{T}}w}\Big) \tag{1}$$

式中:X 为共现矩阵;W 为单词矩阵;C 为上下文矩阵,V 中单词 w、上下文单词 c 对应的词向量构成 W 和 C 的列向量,d_w 为单词 w 所在 X 列的列向量。$S_w = S_{w,1}\times S_{w,2}\times\cdots\times S_{w,c}\times\cdots\times S_{w,|V|}$,表示 $S_{w,c}$ 的笛卡尔乘积,$S_{w,c} = \{0,1,\cdots,k\ \frac{\#(w)\#(c)}{\sum_{w,c\in V}\#(w,c)}\}$ k 是一个参数。

2　融合语义信息的矩阵分解词向量学习模型

2.1　提取语义信息并构建语义矩阵

本文选择 WordNet 做先验知识库。WordNet 是一个覆盖范围较广的英语词汇语义网,它把含义相同的单词组织在同义词集合中,每个同义词集合都代表一个基本的语义概念,并且这些集合之间也由各种关系(例如整体部分关系、上下文关系)连接。

本文基于同义词集合及集合间的关系词构建一个语义关系矩阵 $S\in\mathbb{R}^{V\times V}$,它的每一个元素 $S_{ij} = S(w_i,w_j)$ 表示词汇表 V 中第 i 个单词 w_i 与第 j 个单词 w_j 之间的语义相关性。如果 $S_{ij}=0$ 表示单词 w_i 与 w_j 没有语义相关性,反之 $S_{ij}\neq 0$ 则表示单词 w_i 与 w_j 具有相关性。简单起见,本文将语义关系矩阵 S 构建成 0-1 矩阵,如果单词 w_i 与 w_j 具有上述语义关系则令 $S_{ij}=1$,否则 $S_{ij}=0$。

2.2　构建语义约束模型

本文构建语义约束模型的前提是具有语义相关性的词对 w_i、w_j 学到的词向量更相似,在实数空间中有更近的距离,本文采用向量的欧氏距离作为度量词对相似程度的标尺,即 $d(w_i,w_j) = \|w_i - w_j\|^2$。因此,语义约束模型可以表示为

$$R = \sum_{w_i,w_j\in V} S_{ij}\|w_i - w_j\|^2 =$$

$$\sum_{i,j=1}^{|V|} S_{ij}(w_i^{\mathrm{T}}w_i + w_j^{\mathrm{T}}w_j - 2w_i^{\mathrm{T}}w_j) =$$

$$\sum_{i=1}^{|V|}(\sum_{j=1}^{|V|}S_{ij})w_i^{\mathrm{T}}w_i + \sum_{j=1}^{|V|}(\sum_{i=1}^{|V|}S_{ij})w_j^{\mathrm{T}}w_j - 2\sum_{i,j=1}^{|V|}S_{ij}w_i^{\mathrm{T}}w_j =$$

$$\sum_{i=1}^{|V|}S_i w_i^{\mathrm{T}}w_i + \sum_{j=1}^{|V|}S_j w_j^{\mathrm{T}}w_j - 2\sum_{i,j=1}^{|V|}S_{ij}w_i^{\mathrm{T}}w_j =$$

$$\mathrm{tr}(W^{\mathrm{T}}S_{\mathrm{row}}W) + \mathrm{tr}(W^{\mathrm{T}}S_{\mathrm{col}}W) - 2\mathrm{tr}(W^{\mathrm{T}}SW) =$$

$$\mathrm{tr}(W^{\mathrm{T}}(S_{\mathrm{row}} + S_{\mathrm{col}} - 2S)W)$$

最终所得语义约束模型为

$$R = \mathrm{tr}(W^{\mathrm{T}}(S_{\mathrm{row}} + S_{\mathrm{col}} - 2S)W) \tag{2}$$

式中:tr(·)表示矩阵的迹;S_i 表示语义矩阵 S 第 i 行全部元素值的加和,即 S 的第 i 行和;S_j 表示语义矩阵 S 第 j 列全部元素值的加和,即 S 的第 j 列和;S_{row} 表示以 S_i 为对角元素值的对角矩阵,S_{col} 表示以 S_j 为对角元素值的对角矩阵。

2.3　模型融合

将语义约束模型 R 与 EMF 相结合,得到融合语义信息的矩阵分解词向量学习模型 KbEMF:

$$O = -\mathrm{tr}(X^{\mathrm{T}}C^{\mathrm{T}}W) + \sum_{w\in V}\ln\Big(\sum_{X'_w\in S_w}\mathrm{e}^{X'^{\mathrm{T}}_w C^{\mathrm{T}}w}\Big) + \gamma\mathrm{tr}(W^{\mathrm{T}}(S_{\mathrm{row}} + S_{\mathrm{col}} - 2S)W) \tag{3}$$

式中 γ 是语义组合权重,表示语义约束模型在联合模型中所占的比重大小。γ 在词向量学习过程中扮演相当重要的角色,该参数设置值过小时会弱化先验知识对词向量学习的影响,若过大则会破坏词向量学习的通用性,无论哪种情况都不利于词向量的学习。该模型目标在于最小化目标函数 O,采用变量交替迭代策略求取最优解。当 $\gamma=0$ 时表示没有融合语义信息,即为 EMF 模型。

2.4　模型求解

目标函数,即式(3)不是关于 C 和 W 的联合凸函数,但却是关于 C 或 W 的凸函数,因此本文采用被广泛应用于矩阵分解的变量交替迭代优化策略求取模型的最优解。分别对 C、W 求偏导数,得到

$$\frac{\partial O}{\partial C} = (E_{X|C^{\mathrm{T}}W}X - X)W^{\mathrm{T}} \tag{4}$$

$$\frac{\partial O}{\partial W} = C(E_{X|C^{\mathrm{T}}W}X - X) + \gamma(L + L^{\mathrm{T}})W \tag{5}$$

式中:$L = S_{\mathrm{row}} + S_{\mathrm{col}} - 2S$;$E_{X|C^{\mathrm{T}}W}X$ 位于 w 行 c 列的值是 $E_{X|C^{\mathrm{T}}W}X(w,c) = Q(w,c)\sigma(c^{\mathrm{T}}w)$,$Q$ 位于 w 行 c 列的值是 $Q(w,c) = k\ \frac{\#(w)\#(c)}{\#(w,c)} + \#(w,c)$,$\sigma(x) = \frac{1}{1+\mathrm{e}^{-x}}$。本文的优化更新策略为

$$W \leftarrow W + \eta[C(E_{X|C^{\mathrm{T}}W}X - X) + \gamma(L + L^{\mathrm{T}})W] \tag{6}$$

$$C \leftarrow C + \eta[(E_{X|C^{\mathrm{T}}W}X - X)W^{\mathrm{T}}] \tag{7}$$

在一次循环中先对 W 迭代更新,直到目标函数 O 对 W 收敛为止,然后对 C 迭代更新,再次使目标函数 O 对 C 收敛,至此一次循环结束,依此循环下去直到最终目标函数关于 C 和 W 都收敛为止。

算法 KbEMF 算法的伪代码

输入 共现矩阵 $\boldsymbol{X}$,语义关系矩阵 $\boldsymbol{S}$,学习率 η,最大迭代次数 K,k;

输出 $\boldsymbol{W}_K$,$\boldsymbol{C}_K$。

1)随机初始化: $\boldsymbol{W}_0$,$\boldsymbol{C}_0$

2)for $i = 1$ to K,执行

3)$\boldsymbol{W}_i = \boldsymbol{W}_{i-1}$

4)for $j = 1$ to k, 执行

5) $\boldsymbol{W}_i = \boldsymbol{W}_i + \eta[\boldsymbol{W}_{i-1}(E_{X \mid c_{i-1}^{\mathrm{T}}\boldsymbol{W}_i}\boldsymbol{W} - \boldsymbol{W}) + \gamma(L + L^{\mathrm{T}})\boldsymbol{W}_i]$

6)$j=j+1$

7) $\boldsymbol{C}_i = \boldsymbol{C}_{i-1}$

8)for $j=1$ to k, 执行

9) $\boldsymbol{C}_i = \boldsymbol{C}_i + \eta(E_{X \mid c_i^{\mathrm{T}}\boldsymbol{W}_i}\boldsymbol{X} - \boldsymbol{X})\boldsymbol{W}_i^{\mathrm{T}}$

10)$j=j+1$

11)$i=i+1$

3 实验与结果

本节主要展示融合语义信息后获取的词向量在单词类比推理和单词相似度量任务上的性能表现。首先介绍实验数据集及实验设置,然后分别描述每个实验的任务和结果,并分析实验结果。

3.1 数据集

本实验选择 Enwik9[1] 作为训练语料库,经过去除原始语料库中 HTML 元数据、超链接等预处理操作后,得到一个词汇量将近 13 亿的训练数据集。然后通过设置单词过滤词频限制词汇表的大小,把低于设定过滤词频的单词剔除词汇表,因此,不同过滤词频产生不同大小的词汇表。

本实验选用 WordNet[2] 作为知识库,WordNet[2] 有 120 000同义词集合,其中包含150 000单词。本文借助 JWI[3] 从 WordNet[2] 中抽取单词间的语义关系:同一个同义词集合内单词对的同义关系,以及不同集合间单词对的上下位关系。

不同的实验任务所用的测试数据集也不相同。

在单词类比推理任务中,本文使用的测试集为谷歌查询数据集(Google query dataset[4]),该数据集包含19 544个问题,共 14 种关系类型,其中 5 种语义关系,9 种语法关系。在单词相似度量任务中,本文使用下列 3 个数据集:Luong 等[24] 使用的稀有单词,Finkelstein 等[25] 使用的 Wordsim-353 (WS353) 数据集(RW),Huang 等[6] 发布的上下文单词相似数据集(SCWS)。它们分别包含 2003、2034、353 个单词对及相应的人工标注的相似度分值。

3.2 实验设置

下列实验展示了由 KbEMF 获取的词向量在不同任务中的性能表现。为保持实验效果的一致性,所有模型设置相同的参数。词向量维数统一设置为 200,学习率设置为 6×10^{-7},上下文窗口为 5,迭代次数设置为 300。

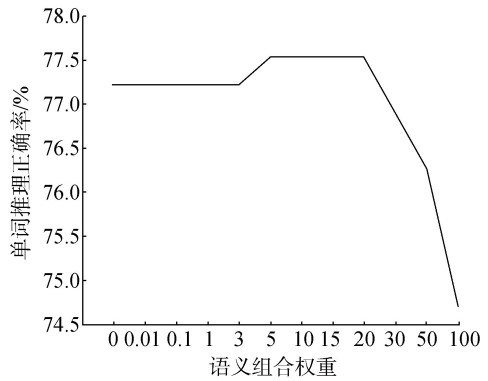
另外,语义组合权重的大小也对实验有重要影响。对于单词类比推理和单词相似度量任务本文均采取相同的实验策略寻找最佳语义组合权重,下面以单词类比推理任务为例详细说明最佳语义组合权重找寻的实验过程。设定 $\gamma \in [0.01, 100]$,首先实验 $\gamma = 0.01, 0.1, 1, 10, 100$ 的单词推理正确率,如图 1 (b)所示, $\gamma = 0.01, 0.1, 1$ 时 KbEMF 没有提升实验效果,因为语义信息所起作用太小;在 $\gamma = 100$ 时 KbEMF 实验效果反而更差,这是过分强调语义信息破坏了词向量的通用性;只有在 $\gamma = 10$ 时 KbEMF 效果较好,则最佳语义组合权重在 $\gamma = 10$ 附近的可能性最大。然后在 $\gamma \in [1, 10]$ 和 $\gamma \in [10, 100]$ 采取同样的策略继续寻找下去,最终会得到最佳组合权重。实验结果表明,不同任务在不同词频下的最优语义组合权重也不同。

3.3 单词类比推理

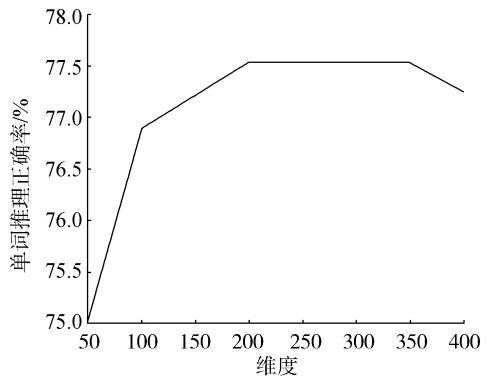
给出一个问题 $a:b::c:d$,a、b、c、d 各表示一个单词其中 d 是未知的,类比推理任务的目标在于找到一个最合适的 d 使得 a, b, c, d 的词向量满足 $\mathrm{vec}(d)$ 与 $\mathrm{vec}(b)-\mathrm{vec}(a)+\mathrm{vec}(c)$ 的余弦距离最近。例如,语义推理 Germary:Berlin::France:d,则需要找出一个向量 $\mathrm{vec}(d)$,使它与 vec(Berlin)-vec(Germary)+vec(France) 最近似,如果 $\mathrm{vec}(d)$ 对应的 d 是 Paris 则推理正确。同理,又如语法推理 quick:quickly::slow:d,如果找到 d 是 slowly 则推理正确。该实验任务的评价指标是推理出单词 d 的正确率,正确率越高,则 KbEMF 学得的词向量越好。

本实验评估了不同参数设置对 KbEMF 模型影响,图 1 是词频为6 000次时,分别改变模型中词向量维度及语义组合权重所绘制的。

从图 1 (a)可以看出,词向量维度小于 200 时,随着词向量维度增加单词推理正确率在提升,词向量维度在 200~350 之间实验效果趋向于稳定,因此在同时兼顾实验速度与效果的情况下,本文选择学习 200 维度的词向量。



(a)向量维度



(b)语义组合权重

图 1　KbEMF 在不同向量维度和语义组合权重的正确率

Fig.1　Performance when incorporating semantic knowledge related to word analogical reasoning for different vector sizes and semantic combination weights

图 1 (b)中随着语义组合权重增大,单词推理正确率在提升,继续增大正确率反而减小,说明过大或过小的语义组合权重都不利于学习词向量。从该实验还可以看出,语义组合权重在[5,20]之间单词推理正确率最高,词向量在该任务中表现最优。

图 2 展示了在不同过滤词频下,KbEMF 的单词推理正确率均在不同程度上高于 EMF,尤其在词频为 3 500 时效果最佳。对于不同词频,该实验均设置语义组合权重 $\gamma = 10$,尽管该参数值在某些词频下不是最优的,却在一定程度上说明本文模型的普遍适用性。

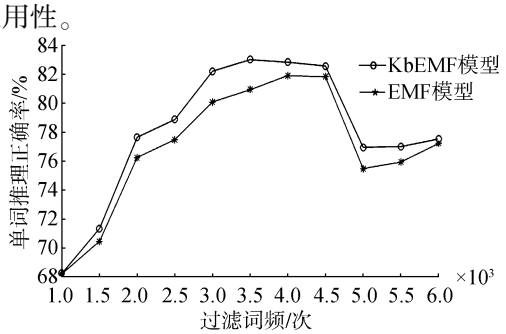


图 2　不同过滤词频下 EMF 与 KbEMF 的正确率对比

Fig.2　Performance of KbEMF compared to EMF for different word frequencies

下面通过将 KbEMF 与 EMF、Retro(CBOW)、Retro(Skip-gram)[5]、SWE 进行比较来说明 KbEMF 的优越性。Retro 根据知识库信息对预先训练好的词向量进行微调,该模型的缺点在于无法在语料库学习词向量阶段利用丰富的语义信息。虽然 SWE 同时利用了语义信息和语料库信息学习词向量,但该模型的基础框架 skip-gram 只考虑了语料库的局部共现信息。本文提出的 KbEMF 则克服了上述模型的弱点,同时利用语料信息和语义信息学习词向量,并且它所分解的共现矩阵覆盖了语料库的全局共现信息。表 1 展示了词频为 3 500 时 KbEMF 与 EMF、Retro(CBOW)、Retro(Skip-gram)[5]、SWE 的单词推理正确率。

表 1　KbEMF 与其他方法的单词推理正确率

Table 1　Performance of KbEMF compared to other approaches

%

方法	单词推理正确率
EMF	80.93
Retro(CBOW)	78.47
Retro(Skip-gram)	77.30
SWE	79.77
KbEMF	83.01

表 1 中 KbEMF 对应的单词推理正确率最高,这说明该模型所获取的词向量质量最优。

3.4　单词相似度量

单词相似度量是评估词向量优劣的又一经典实验。该实验把人工标注的词对相似度作为词对相似度的标准值,把计算得到的词对向量余弦值作为词对相似度的估计值,然后计算词对相似度的标准值与估计值之间的斯皮尔曼相关系数(spearman correlation coefficient),并将它作为词向量优劣的评价指标。斯皮尔曼相关系数的值越高表明单词对相似度的估计值与标准值越一致,学习的词向量越好。

由于单词相似度量希望相似度高或相关度高的词对间彼此更靠近,语义信息的融入使具有强语义关系的词对获得更相似的词向量。那么计算所得的关系词对向量的余弦值越大,词对相似度的标准值与估计值之间的斯皮尔曼相关系数就越高。

与单词类比推理实验过程类似,通过调整 KbEMF 模型参数(词向量维度、语义组合权重以及单词过滤词频),获得单词相似度量实验中表现优异的词向量。

本实验比较了 KbEMF 与 SWE、Retro 在单词相似度量任务中的性能表现,结果展示在表 2 中。由

于不同数据集下最佳语义组合权重不同,该实验针对数据集 WS353/SCWS/RW 分别设置语义组合权重为 $\gamma=1,\gamma=1,\gamma=15$。

表 2 不同数据集下 KbEMF 与其他方法的斯皮尔曼相关系数

Table 2 Spearman correlation coefficients of KbEMF compared to other approaches on different datasets

方法	数据集		
	WS353	SCWS	RW
EMF	0.791 8	0.647 4	0.678 6
Retro(CBOW)	0.781 6	0.668 5	0.607 1
Retro(Skip-gram)	0.693 0	0.644 9	0.714 3
SWE	0.796 5	0.659 3	0.642 9
KbEMF	0.799 9	0.674 0	0.750 0

表 2 中 KbEMF 在上述 3 个数据集的斯皮尔曼相关系数均有所提升,因为 KbEMF 相比较 Retro 在语料库学习词向量阶段就融入了语义知识库信息,相较于 SWE 则运用了语料库全局的共现信息,因此表现最好。尤其 KbEMF 在 RW 上的斯皮尔曼相关系数提升显著,这说明语义知识库信息的融入有助于改善学习稀有单词的词向量。

4 结束语

学习高效的词向量对自然语言处理至关重要。仅依赖语料库学习词向量无法很好地体现单词本身的含义及单词间复杂的关系,因此本文通过从丰富的知识库提取有价值的语义信息作为对单一依赖语料库信息的约束监督,提出了融合语义信息的矩阵分解词向量学习模型,该模型大大改善了词向量的质量。在实验中将 Enwik9 作为训练文本语料库并且将 WordNet 作为先验知识库,将学到的词向量用于单词相似度量和单词类比推理两项任务中,充分展示了本文模型的优越性。

在后续的研究工作中,我们将继续探索结合其他知识库(如 PPDB、WAN 等),从中抽取更多类型的语义信息(如部分整体关系、多义词等),进而定义不同更有针对性的语义约束模型,进一步改善词向量。并将它们用于文本挖掘和自然语言处理任务中。

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, 2010: 384-394.

[2] LIU Y, LIU Z, CHUA T S, et al. Topical word embeddings [C]//Association for the Advancement of Artificial Intelligence. Austin Texas, USA, 2015: 2418-2424.

[3] MAAS A L, DALY R E, PHAM P T, et al. Learning word vectors for sentiment analysis[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland Oregon, USA, 2011: 142-150.

[4] DHILLON P, FOSTER D P, UNGAR L H. Multi-view learning of word embeddings via cca[C]//Advances in Neural Information Processing Systems. Granada, Spain, 2011: 199-207.

[5] BANSAL M, GIMPEL K, LIVESCU K. Tailoring continuous word representations for dependency parsing[C]//Meeting of the Association for Computational Linguistics. Baltimore Maryland, USA, 2014: 809-815.

[6] HUANG E H, SOCHER R, MANNING C D, et al. Improving word representations via global context and multiple word prototypes[C]//Meeting of the Association for Computational Linguistics. Jeju Island, Korea, 2012: 873-882.

[7] MNIH A, HINTON G. Three new graphical models for statistical language modelling[C]//Proceedings of the 24th International Conference on Machine Learning. New York, USA, 2007: 641-648.

[8] MNIH A, HINTON G. A scalable hierarchical distributed language model [C]//Advances in Neural Information Processing Systems. Vancouver, Canada, 2008:1081-1088.

[9] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model[J]. Journal of machine learning research, 2003, 3(02): 1137-1155.

[10] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12(8): 2493-2537.

[11] MIKOLOV T, CHEN K, CORRADO G, ET AL. Efficient estimation of word representations in vector space[C]// International Conference on Learning Representations. Scottsdale, USA,2013.

[12] BAIN J, Gao B, Liu T Y. Knowledge-powered deep learning for word embedding[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2014: 132-148.

[13] LI Y, XU L, TIAN F, ET AL. Word embedding revisited: a new representation learning and explicit matrix factorization perspective [C]//International Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 3650-3656.

[14] LEVY O, GOLDBERG Y. Neural word embedding as implicit matrix factorization [C]//Advances in Neural Information Processing Systems. Montreal Quebec, Canada, 2014: 2177-2185.

[15] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//Conference on Empirical Methods in Natural Language Processing. Doha,

Qatar, 2014: 1532-1543.

[16] BIAN J, GAO B, LIU T Y. Knowledge-powered deep learning for word embedding [C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin, Germany, 2014: 132-148.

[17] XU C, BAI Y, BIAN J, et al. Rc-net: a general framework for incorporating knowledge into word representations [C]//Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. Shanghai, China, 2014: 1219-1228.

[18] YU M, DREDZE M. Improving lexical embeddings with semantic knowledge[C]//Meeting of the Association for Computational Linguistics. Baltimore Maryland, USA, 2014: 545-550.

[19] LIU Q, JIANG H, WEI S, et al. Learning semantic word embeddings based on ordinal knowledge constraints[C]// The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing. Beijing, China, 2015: 1501-1511.

[20] FARUQUI M, DODGE J, JAUHAR S K, et al. Retrofitting word vectors to semantic lexicons[C]//The 2015 Conference of the North American Chapter of the Association for Computational Linguistics. Colorado, USA, 2015: 1606-1615.

[21] LEE D D, SEUNG H S. Algorithms for non-negative matrix factorization [C]//Advances in Neural Information Processing Systems.Vancouver, Canada, 2001: 556-562.

[22] MNIH A, SALAKHUTDINOV R. Probabilistic matrix factorization [C]//Advances in Neural Information Processing Systems. Vancouver, Canada, 2008: 1257-1264.

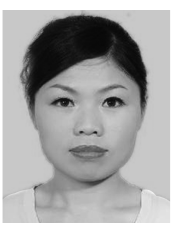
[23] SREBRO N, RENNIE J D M, JAAKKOLA T. Maximum-margin matrix factorization [J]. Advances in neural information processing systems, 2004, 37(2):1329-1336.

[24] LUONG T, SOCHER R, MANNING C D. Better word representations with recursive neural networks for morphology [C]//Seventeenth Conference on Computational Natural Language Learning. Sofia, Bulgaria,2013: 104-113.

[25] FINKELSTEIN R L. Placing search in context:the concept revisited[J]. ACM transactions on information systems, 2002, 20(1): 116-131.

作者简介:

陈培,女,1990 年生,硕士研究生,主要研究方向为自然语言处理、情感分析。



景丽萍,女,1978 年生,教授,博士,主要研究方向为数据挖掘、文本挖掘、生物信息学、企业智能。