分布式单词表示综述

孙 飞[1),2),3)] 郭嘉丰[1),2)] 兰艳艳[1),2)] 徐 君[1),2)] 程学旗[1),2)]

[1)](中国科学院网络数据科学与技术重点实验室 北京 100190)

[2)](中国科学院计算技术研究所 北京 100190)

[3)](中国科学院大学 北京 100190)

摘 要 单词表示作为自然语言处理的基本问题,一直广受关注.传统的独热表示丢失了单词间的语义关联,因而在实际使用中易受数据稀疏问题困扰.而分布式表示通过将单词表示为低维稠密实数向量,捕捉单词间的关联信息.该表示方式可在低维空间中高效计算单词间的语义关联,有效解决数据稀疏问题.作为神经网络模型的基本输入,单词分布式表示伴随着深度学习被广泛应用于自然语言处理领域的方方面面.从早期的隐式语义分析,到最近的神经网络模型,研究人员提出了各种各样的模型来学习单词的分布式表示.本文梳理了单词分布式表示学习的发展脉络,并从模型利用上下文入手,将这些模型统一在分布语义假设框架下,它们的区别只在于建模了单词不同的上下文.以隐式语义分析为代表的话题模型,利用文档作为上下文,建模了单词间的横向组合关系;以神经网络语言模型为代表的工作,则利用单词周围单词作为上下文,建模了单词间的纵向聚合关系.此外,本文还总结了单词分布式表示目前面临的主要挑战,包括多义词的表示、稀缺单词表示学习、细粒度语义建模、单词表示的解释性以及单词表示的评价,并介绍了最新的已有解决方案.最后,本文展望了单词表示未来的发展方向与前景.

关键词 单词表示;分布式表示;分布式单词表示;表示学习;深度学习

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2019.01605

A Survey on Distributed Word Representation

SUN Fei[1),2),3)] GUO Jia-Feng[1),2)] LAN Yan-Yan[1),2)] XU Jun[1),2)] CHENG Xue-Qi[1),2)]

[1)](*CAS Key Lab of Network Data Science and Technology, Beijing* 100190)

[2)](*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[3)](*University of Chinese Academy of Sciences, Beijing* 100190)

Abstract As a fundamental problem in natural language processing, word representation is always widely concerned by the society. Traditional one-hot representations suffer from the data sparsity in practice due to missing semantic relation between words. Different from the one-hot representations, distributed word representations encode the semantic meaning of words as dense, real-valued vectors in a low-dimensional space. As a result, the distributed word representations can alleviate the data sparsity isses. As the inputs of neural network models, distributed word representations have been widely used in natural language processing along with deep learning. From latent semantic indexing to neural language model, researchers have developed various methods to learn distributed word representations. In this paper, we comb the development of models for learning distributed word representations. Furthermore, we find that all these models are built

收稿日期:2016-05-04;在线出版日期:2016-09-21.本课题得到国家"九七三"重点基础研究发展规划项目基金(2014CB340401,2013CB329606)、国家自然科学基金(61232010,61472401,61425016,61203298)、中国科学院青年创新促进会(20144310,2016102)资助.**孙 飞**,博士研究生,主要研究领域为文本表示学习与文本挖掘.E-mail:ofey.sunfei@gmail.com.**郭嘉丰**,博士,副研究员,中国计算机学会(CCF)会员,主要研究领域为信息检索与数据挖掘.**兰艳艳**,博士,副研究员,中国计算机学会(CCF)会员,主要研究领域为统计机器学习、排序学习和信息检索.**徐 君**,博士,研究员,中国计算机学会(CCF)会员,主要研究领域为信息检索与数据挖掘.**程学旗**,博士,研究员,中国计算机学会(CCF)会员,主要研究领域为网络科学、互联网搜索与挖掘和信息安全等.

on the distributional hypothesis but with different contexts. From this perspective, we can group these models into two classes, syntagmatic and paradigmatic. Models like latent semantic indexing using documents as the contexts for words to capture the syntagmatic relations between words. While, models like neural language models capture the paradigmatic relations between words by the contexts surrounding the words. Then, we summarize the key challenges and the latest solutions, like representations for polysemous words and rare words, fine-grained semantic modeling, interpretability for distributed word representations, and evaluation for word representation. At last, we give a future outlook on the research and application directions.

Keywords word representation; distributed representation; distributed word representation; representation learning; deep learning

1 引 言

表示学习作为机器学习的一个基本问题,其结果直接影响着整个机器学习系统的性能[1].单词作为语言的基本单元,其表示学习也一直是文本处理领域的核心问题.

长久以来,自然语言处理(Natural Language Processing, NLP)等领域最常用的单词表示方法一直是独热表示(One-Hot Representation).该方法仅仅将单词离散符号化,将单词表示为一个只有某一维非 0 的向量,且每个单词使用不同维度.显而易见,独热表示存在着诸多问题.首先,这种表示不包含任何语义信息,无法表示单词间的语义关联差异.比如,相比于"机器","猫"与"狗"语义更相似.此外,为表示 N 个单词,独热表示需要的表示长度为 N,这样才能区分这 N 个单词.正因此,独热表示在实际应用中往往也会面临着参数的组合爆炸以及数据稀疏问题.以语言模型(Language Modeling,LM)为例,使用独热表示,即使是简单的三元语言模型,其参数空间也远超普通计算机的能力,且伴随着严重的稀疏问题.

近年来,深度学习广泛应用于自然语言处理领域各方面,并取得了良好结果.然而,正如 Manning 所言[2],深度学习目前在自然语言处理领域,并没有取得如语音和图像领域一样的突破,而其取得的提升,更多地来自于单词的分布式表示(Distributed Representation)——将单词表示为低维实数向量.因而伴随着深度学习在自然语言处理领域的火热,单词分布式表示也同样获得了广泛关注.

与独热表示只使用向量的一个维度不同,单词的分布式表示,使用低维稠密实数向量来表示单词.在该低维向量空间中,可以方便地根据距离或角度等度量方式,衡量两个单词间的相似程度.Bengio 等人[3]将分布式表示应用于单词,结合神经网络,训练语言模型,成功解决了传统概率语言模型中的参数组合爆炸以及稀疏问题.

然而,早期的单词分布式表示学习模型,计算复杂难堪实际应用.

随着以 Word2Vec[4]为代表的单词分布式表示学习模型的提出,使得快速地从大量无标注的自然语言文本中自动学习得到单词的表达成为现实.这种表示可以直接集成于现有机器学习系统,在近年已被广泛应用于自然语言处理的各方面,如情感分析[5-6]、句法分析[7-8]、词性标注[9]、机器翻译[10-12]、话题模型[13-15]等.

近年来,单词分布式表示研究取得了众多进展,大量的单词表示学习模型被提出,同时这些表示也被广泛应用于自然语言处理的方方面面并取得显著性能提升.然而,在引起广泛关注和研究兴趣的同时,该方向也面临着诸多挑战.本文将介绍单词分布式表示学习的最新进展,总结该领域面临的主要挑战和已有解决方案,并展望未来发展方向与前景.

2 单词表示形式简介

介绍单词分布式表示学习的主要模型与挑战前,本节首先介绍单词表示的基本概念与理论.

2.1 独热表示

在分布式表示之前,有着更简单直接的单词表示方法,即局域性表示(Local Representation).该方法同样将单词表示为一个向量,但是对于每个单词,其只使用向量中互不相交的维度来表示.极端情况下,只使用一个维度,则被称为独热表示.如

“狗”表示为[0,0,…,0,1,0,0,…,0]；

“猫”表示为[0,0,…,0,0,1,0,…,0]；

“机器”表示为[0,0,…,1,0,0,0,…,0].

独热表示假设所有单词都是相互独立无关的.在其表示空间中,所有的单词向量都是正交的.在此情形下,如通过余弦距离度量单词相似度,则得到的单词间相似度均为 0;若使用欧式距离度量单词相似度,则所有单词间语义相似度均相等.无论是“狗”和“猫”这样的语义相近的单词,还是“狗”和“机器”这样无关的词,它们之间的距离都是一样的.这种表示方式,丢失了单词之间的语义相关信息.这也正是独热表示以及以其为基础的词袋模型(Bag of Words, BoW)容易受数据稀疏问题影响的根本原因.

此外,独热表示在实际应用时经常会面临着维度灾难问题.以概率语言模型(Probabilistic Language Modeling)为例,假设单词的集合为 V,那么即使是简单的三元语言模型,其参数空间大小则为 $|V|^3$.

假设词表中只 10 万单词,则其参数空间为 10^{15},已经远超普通计算机的计算能力.与此同时,它也面临严重的数据稀疏问题.如大部分三元组(trigram)的参数取值都是 0,而当在测试时遇到了在训练集中没有出现的三元组,模型便无法简单处理,必须借助于复杂的平滑策略.

当然,独热表示也并非一无是处.与分布式表示相比,独热表示无需学习过程,简单高效.配合最大熵(Maximum Entropy)、支持向量机(Support Vector Machine)、条件随机场(Conditional Random Field)等学习算法,独热表示在文本分类、文本聚类、词性标注等众多问题上都取得良好结果.因此,长久以来它被广泛应用于自然语言处理、信息检索等领域.此外,相比分布式表示,独热表示具有更强的判别能力.这是因为分布式表示将单词表示为低维稠密实数向量后,语义相似的单词在向量表示形式下变的非常接近而难以区分.而独热表示则不会有这个问题.这使得独热表示在文本分类这类需要很强判别能力的任务上依然是一个很强的基准模型[16-17].如在传统的文本分类领域,基于独热表示使用 TF-IDF 权重的词袋模型依然是一个很强的基准模型[18].而目前,基于单词分布式表示的深度学习模型,也只是在情感分类领域相对独热表示具有明显优势[5].此外,对于 ad-hoc 检索,这种关键词匹配占主导作用的应用场景,基于独热表示的词袋模型目前依然是主流选择[19].

2.2　分布式表示

分布式表示的概念最早由 Hinton 等人[20]区别于独热表示提出,用以表示概念.这种表示方式的思想来源于认知表示,一个对象可以通过刻画它的各种属性来高效表示(所有的属性状态,都可以为激活或者非激活).而这些属性,又同时与多个概念相关联.这样,一个概念可以通过这些基本属性的激活状态来高效表示.形式化地,与独热表示只使用向量的一个维度不同,分布式表示则是用稠密实数向量来表示一个单词(向量多于一个维度非 0,通常为低维向量).如同样是上文独热表示中示例的 3 个单词,用分布式表示则可能为:

“狗”表示为[0.14,…,0.61,…,−0.27]；

“猫”表示为[0.18,…,0.71,…,−0.31]；

“机器”表示为[−0.43,…,0.02,…,0.97].

相比于独热表示,分布式表达具有许多根本上的优点.首先,分布式表示可以编码不同单词之间的语义关联.如上例中,分布式表示可以让“狗”和“猫”在大多数维度上相近,而只在少数表征各自不同属性(如习性)的维度上取值不一致,这样“狗”和“猫”的向量之间的距离可以远小于“狗”和“机器”之间的距离.随之而来的是,分布式表示具有更强的泛化能力.当学习到已有概念 a 的新知识时,使用分布式表示,可以自动的泛化到相似的概念 b.假设对于概念 a,学习到新知识 a 喜欢概念 c,表示为 like(a,c),则此知识可以自动泛化到 like(b,c),因为 a 与 b 具有相似的表达输入.再者,分布式表示具有比独热表示更强的表示能力.即使只使用二值表示(每一维取值只能为 0 或 1),长度为 n 的独热表示只能表示 n 个不同概念,而分布式表达则可以 2^n 个不同概念.

因为以上优点,在实际应用中,分布式表示能有效缓解数据稀疏问题.以语言模型为例,一方面使用分布式表示只需要远少于独热表示的参数复杂度;另一方面对于训练集中没有出现的序列,分布式表示可以利用训练集中相似的序列来帮助估计此序列的概率[3].因其能够充分利用对象间的语义关联,缓解数据稀疏问题,分布式表示自提出以来,已被广泛应用于单词[3-4,21-23]、短语[7,24-28]、概念[29]、句子[30-31]、文档[32]和社会网络[33-34]等对象的表示学习中.本文主要以单词为主体,讨论分布式表示在单词表示学习中的研究进展.

需要注意的是,在自然语言处理领域,还有一个容易与分布式表示混淆的概念:分布表示(Distributional Representation)或分布语义(Distributional Semantic).分布语义主要是指单词的语义是通过上下文的分布来表示,一般表示形式是高维向量,每个维度对应于与单词的上下文.这与分布式表示是两

个层面的概念.分布语义强调单词的语义来自于其上下文,这里的“分布(Distributional)”指的是概率分布.而“分布式(Distributed)”表示,则是强调表示形式采用的是一种分布式的形式.事实上分布式表示,同样也多是用分布语义来学习低维表示.而本文关注于低维实数向量的分布式表示.在不做特殊说明的情况下,下文所提“表示”、“向量表示”、“词嵌入”,均指分布式表示.

3 单词分布式表示学习主要方法

单词的分布式表示是近年来的研究热点.研究者们提出了多种模型来学习单词的分布式表示.本文将介绍其中几种代表性的方法,以阐述单词分布式表示学习的发展脉络.模型间发展脉络示意图如图1所示.

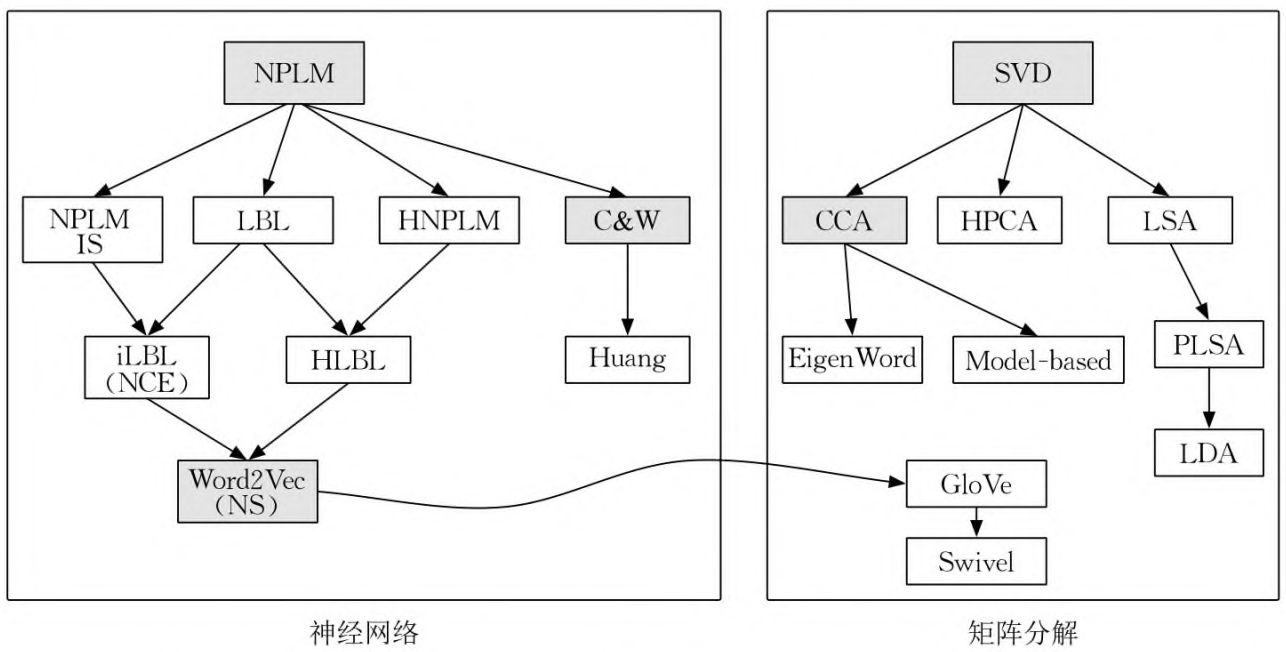

图1 单词表示学习代表性方法发展示意图(方框节点为代表性模型,箭头所示为模型影响方向,加灰节点为本节重点介绍模型,因其衍生出大量后续工作.注意,此图中只列出了代表性工作,省略了大量的细节工作,如C&W、Word2Vec、CCA等,都有大量衍生工作)

3.1 符号说明

为了介绍这些方法,这里首先定义需使用符号,便于下文统一使用.首先,将语料定义为单词序列$[w_1,\cdots,w_N]$,其中$w_i\in V$表示在位置i上的单词.$[c_{i-k},\cdots,c_{i-1},c_{i+1},\cdots,c_{i+k}]$表示单词$w_i$前后长度为$k$范围内的上下文,其中$c_j$表示在位置$j$上的上下文单词,注意这里的$c$可能与$w$是一样的单词,用不同符号只是示意其所处角色不同.本文使用$\boldsymbol{w}_i$与$\boldsymbol{c}_j$分别表示单词w_i和c_j对应的向量表示.下文,如无特殊说明,普通变量x所对应的黑体$\boldsymbol{x}$代表其向量表示.

3.2 神经网络语言模型

学习单词的分布式表示并应用于自然语言处理,可以追溯到1991年Miikkulainen和Dyer[35]的工作,他们尝试使用PDP(Parallel Distributed Processing[36])网络以及分布式表达学习句子中单词所起的角色.

而真正使单词分布式表达获得广泛关注的,则是Bengio等人[3,37]的神经网络概率语言模型(Neural Probabilistic Language Model,NPLM)工作.虽然,Xu和Rudnicky[38]在Bengio之前已使用神经网络学习语言模型.但是他们的网络并没有隐层,而且只使用一个单词作为输入,只建模了二元语言模型.更重要的是,Bengio等人[3,37]的工作提出了一个通用的框架学习单词的分布式表达以及任意的N元语言模型.

而后,伴随着深度学习(Deep Learning,DL)[39]的火热,以及其在语言模型上取得的成功,单词分布式表达同样获得了广泛的关注与研究.而基于神经网络学习单词分布式表达的工作,大体都可追溯到Bengio等人[3]的工作.

概率语言模型,简单来说就是建模单词序列$[w_1,\cdots,w_N]$的概率,可定义为

$$p(w_i,\cdots,w_n)=\prod_{i=1}^{n}p(w_i\mid w_1,,\cdots,w_{i-1}) \quad (1)$$

给定一句话或一段文本,语言模型可以计算它的概率,也因此它已被广泛应用于语音识别(Speech Recognition)[40]、信息检索(Information Retrieval)[41]等领域.从式(1)可以看出,语言模型的关键在于估计给定前文$[w_i,\cdots,w_{i-1}]$后单词w_i的概率.

Bengio等人使用1个3层的神经网络来构建语言模型,其框架如图2所示.

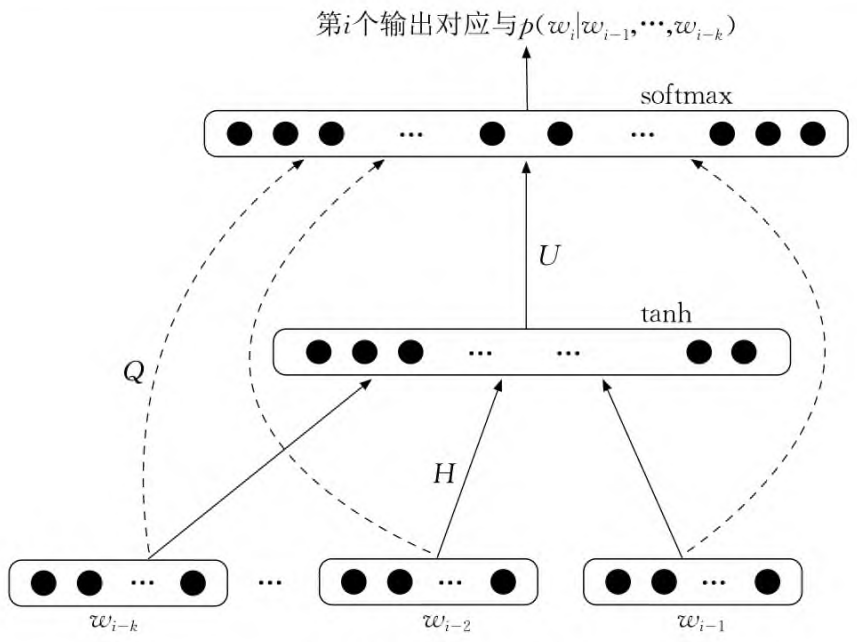

图 2　神经网络语言模型

其根据当前单词 w_i 的前 k 个单词$[w_{i-k},\cdots,w_{i-1}]$的向量输入，经过中间隐层转换后，利用输出层 softmax 层给出在此前文下，任何一个单词的概率：

$$p(w_i|w_{i-1},\cdots,w_{i-k})=\frac{\exp(\boldsymbol{y}_{w_i})}{\sum_{w_j}\exp(\boldsymbol{y}_{w_j})}$$

$$\boldsymbol{y}=\boldsymbol{b}+\boldsymbol{Q}\boldsymbol{x}+\boldsymbol{U}\tanh(\boldsymbol{d}+\boldsymbol{H}\boldsymbol{x})$$

$$\boldsymbol{x}=(\boldsymbol{w}_{i-k},\cdots,\boldsymbol{w}_{i-1})$$

其中，$\boldsymbol{Q},\boldsymbol{U},\boldsymbol{H}$ 对应神经网络中边权重；$\boldsymbol{b},\boldsymbol{d}$ 为网络的偏差，$\boldsymbol{y}$ 是一个长度为$|V|$的向量. 以极大似然为目标函数，使用随机梯度下降，便可以学习更新网络中所有的参数，包括单词的分布式表示以及语言模型的参数.

NPLM 通过将词表示为分布式表达，有效的避免了维度灾难的问题，同时编码了词和词之间的联系，因而自带平滑效果，无需传统 n 元语言模型中那些复杂的平滑算法，如 Modified Kneser-Ney[42].

在 Brown 与 APNews 数据集上做的对比实验表明，NPLM 模型效果都要显著好于精心设计平滑算法的 n 元语言模型[42-43].

问题在于，为什么这样一个神经网络在建模语言模型的时候，能学习出单词的分布式表示？为什么得到的向量表示能捕获到单词间的语义关联？

其核心思想在于，对于同一个单词，在其前面出现的上下文单词总是相似的. 对于 NPLM 来说，也就是相似的输出，需要相似的输入，而这里输入便是单词的表达.

然而，此模型在计算方面存在明显不足. NPLM 使用 softmax 层估计下一个词为 w_i 的概率，而这一层的维度是词表大小，分母需要进行$|V|$次计算. 因而导致学习与推断的过程都异常耗时.

Bengio 等人在文献[3]中表示，在只有13 994 528个单词的 APNews 数据集上使用 40 个 CPU 只训练 5 轮，就已耗费 3 周时间，可见 NPLM 的复杂度过高，难以应用于大规模数据.

3.2.1　神经网络语言模型的加速

由于神经网络语言模型效率低下，难以实际使用，因此早期使用神经网络语言模型学习单词表示的工作，主要集中于加速神经网络语言模型. 对于神经网络语言模型加速的工作，主要集中于两方面：(1) 直接近似优化原始目标函数；(2) 简化网络结构. 下文从这两个角度分别简述代表性工作.

(1) 近似优化目标函数

早先，Bengio 与 Ducharme[44] 中提出使用重要性采样(Importance Sampling)的方法近似目标函数梯度中的期望项，使得训练速度提升 100 倍，然而预测代价依然很高.

由于重要性采样的稳定性问题，Minh 等人[45-46]引入噪声对比估计(Noise-Contrastive Estimation, NCE[47])取代重要性采样进行训练过程中的概率估计. 其基本思想在于训练一个使用相同参数的逻辑斯蒂回归(Logistic Regression, LR)，将真实分布的样本从噪声分布中区分出来.

(2) 修改网络结构

Morin 与 Bengio[48] 将 Goodman[49] 用于加速最大熵语言模型的方法应用于神经网络语言模型，提出层次化的神经网络语言模型. 他们将原本 NPLM 中扁平的 softmax 输出层，改为树状输出. 也就是将原本在所有单词上的多分类问题，转换为一系列的二分类问题. 举例来讲，假设原本需要预测的下一个单词 w_i 是“狗”，而现在首先预测它是不是动物，然后预测它是不是哺乳动物，直到最后才预测它是不是狗. 这样，相比原来$|V|$次的指数运算，现在只需要 $\log|V|$次. 在此方式下，输出端树的构建，是一个关键步骤，其直接影响最终的结果. 比如，在文献[48]中虽然速度得到了极大提升，但是结果却有所下降.

随后，Mnih 和 Hinton[23] 提出了更简单直接的 Log-bilinear(LBL)模型，去除了之前 Bengio 等人[3]模型中隐层的非线性计算部分，输入单词的表示经过简单的线性变换后直接与被预测词的向量作交互. 在 LBL 中，上下文单词与目标单词使用的是同一种表示，而在 Mnih 等人的后续工作中[46]则使用了两种不同的表达. 更进一步地，Mnih 和 Hinton[50]也将层次化的思想用于加速 LBL 模型的训练，并取

得了良好的结果.值得注意的是,在这里 Mnih 等人已经开始处理多义词,遗憾的是他们并没有学习单词的多义表示.

3.3 排序模型

前文所述工作,都是通过训练语言模型习得单词的向量表示.而 Mikolov 等人[51]发现,将单词表达学习和语言模型的训练分离,首先使用简单模型在更大语料上学习单词表达,然后以此训练语言模型,同样可以取得很好的效果.此外,研究人员也逐渐尝试将单词的分布式表达应用于除语言模型以外的自然语言处理任务中.如 Collobert 和 Weston[9]提出的 C&W 模型,使用多任务学习(Multitask Learning)学习单词表达,并用于词性标注、命名实体识别、语义角色标注等任务.

对于 C&W 模型,其目标并不是训练语言模型,而是学出好的表达,并应用于多个任务.

因此,它相比传统神经网络语言模型学习单词表示模型主要有两点改进:

(1)C&W 同时使用了单词前后的上下文.这点也成为后来学习单词表示工作的基本做法.

(2)C&W 对单词序列打分使用了排序损失函数,而非基于概率的极大似然估计.其损失函数如下

$$\max(0,1-s(w,c)+s(\tilde{w},c))$$

其中:c 代表单词 w 的上下文;$\tilde{w}$ 表示将当前上下文 c 中的单词 w 替换为了一个随机采样出的无关单词 $\tilde{w}$;s 代表打分函数.打分高,说明这段文本是正确的;打分低,则说明这段文本不合理.显然,在大多数情况下,将普通短语中的特定单词随机地替换为任意单词,得到的都是不正确的短语.因此,模型的目标便是尽量使正确的语言(也就是观测的语料)得分比随机生成的语言的分数高于 1.

在 C&W 模型之前,Okanohara 和 Tsujii[52]已经在语言模型中使用负样例技术.而 C&W 是最早成功将其应用于深度模型中的工作,并启发了后续一系列的基于正负样例排序误差框架的单词表示学习工作,如 Huang 等人[53]多义词表达工作、Luong 等人[54]关于稀缺词表达的工作.同时 C&W 也是早期将卷积神经网络(Convolutional Neural Network,CNN)应用于 NLP 的代表作之一.

而近期,Ji 等人[55]则进一步地将排序损失函数,单词表示学习问题建模为一个排序问题,使用鲁棒排序(Robust Ranking)学习单词表示.在此模型中,折损累积增益(Discounted Cumulative Gain,DCG)式的损失函数自带了关注机制(attention)以及对于噪声的鲁棒.

Lazaridou 等人[56]表明,使用此类排序损失函数,可以解决单词表示空间中的中枢节点(Hubness)问题(指空间中离的很近的点,往往很难区分).

3.4 上下文单词预测模型

前文所述工作表明,使用更简单的网络模型,利用单词前后的上下文,在更大的数据上,可以得到更好的单词表达.据此,Mikolov 等人[4,23]简化了以往的神经网络语言模型,去除了 NPLM 中间的非线性隐藏层,提出两个简单的神经网络模型(Continuous Bag-of-Words,CBOW 和 Skip Gram,SG)来学习单词分布式表示,其框架如图 3 所示.

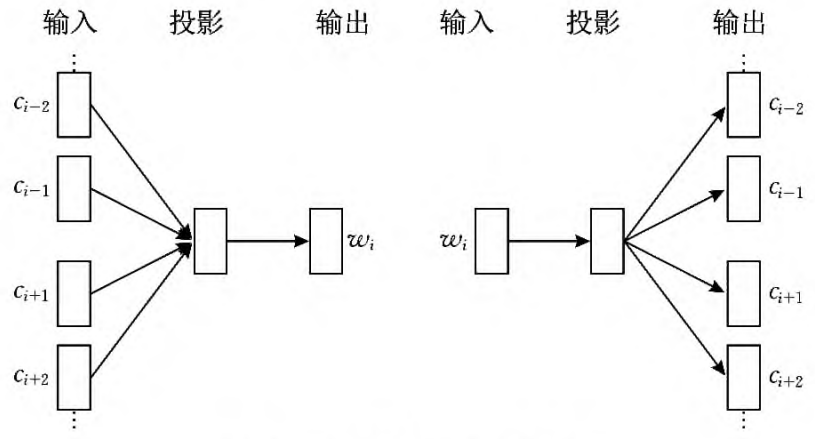

图 3 CBOW 与 SG 框架图

与以往的神经网络模型不同,CBOW 模型去除了非线性隐层,将当前单词 w_i 上下文的表示求和或平均后,直接预测单词 w_i.而 SG 模型则与 CBOW 对称,使用当前单词 w_i 预测其前后上下文中的每一个单词.去掉隐藏层之后,这两个模型从神经网络结构简化为对数线性模型.

CBOW 模型,对于单词 w_i 和其上下文,其目标函数为

$$p(w_i|c)=\frac{\exp(\boldsymbol{w}_i\cdot\boldsymbol{h}_i)}{\sum_{w'}\exp(\boldsymbol{w}'\cdot\boldsymbol{h}_i)}$$

其中 $\boldsymbol{h}_i$ 为单词 w_i 前后上下文单词向量的均值向量.

SG 模型,对于单词 w_i 和其上下文,其目标函数为

$$p(w_i|c)=\prod_{c_j}\frac{\exp(\boldsymbol{w}_i\cdot\boldsymbol{c}_j)}{\sum_{w'}\exp(\boldsymbol{w}'\cdot\boldsymbol{c}_j)}$$

两个模型的求解都是在整个语料上做对数似然最大化.然而,原始目标函数存在着与神经网络语言模型一样的计算问题,分母计算复杂.Mikolov 等人在文献[27]中针对 SG 模型提出了两种优化方法,其中层次 softmax 与前文所述 Morin 等人[48]以及 Mnih 等人[23]的工作相似,此处不再赘述.

除此之外,Mikolov 等人还提出了负采样(Nega-

tive Sampling，NS)技术学习单词表示. 由于 CBOW 与 SG 的目标并不是学习语言模型，因此并不需要像 NCE 那样近似最大化 softmax 的对数概率. Mikolov 转而建模单词与上下文在数据中共现与否的概率，定义如下

$$p(D=1|w_i,c_j)=\frac{1}{1+\exp(-\boldsymbol{w}_i\cdot\boldsymbol{c}_j)}$$

$$p(D=0|w_i,c_j)=\frac{1}{1+\exp(\boldsymbol{w}_i\cdot\boldsymbol{c}_j)}$$

其中：$p(D=1|w_i,c_j)$代表单词 w_i 与上下文 c_j 在数据中共现的概率；$p(D=0|w_i,c_j)$表示单词 w_i 与上下文 c_j 在数据中不共现的概率. 负采样技术便是随机生成一些单词作为当前单词的错误上下文，也就是负样例. 而原本数据中单词和其上下文则是正样本. 最终单词的表示，便是通过极大数据的对数似然求解. 在此情况下，使用了负采样的 SG 模型，对于单词 w_i 以及其上下文 c_j，实际上定义了一个新的目标函数，如下

$$\log\sigma(\boldsymbol{w}_i\cdot\boldsymbol{c}_j)+k\boldsymbol{E}_{\tilde{c}\sim P_{\tilde{C}}}\log\sigma(-\boldsymbol{w}_i\cdot\tilde{\boldsymbol{c}})$$

其中：$\sigma(x)=1/(1+\exp(-x))$；$P_{\tilde{C}}$代表负样例的分布，通常设为词频比例的 3/4 次方，$P_{\tilde{C}}(w)\propto(\#w)^{3/4}$.

实验结果显示，在包含 60 亿单词的 Google news 数据集上进行训练得到的单词表达远超传统的语言模型得到的表达，而训练时间却只有后者的十分之一左右. 实际上，此前从没有模型在如此量级的数据上成功学习出单词表示.

由于 CBOW 和 SG 简单有效，CBOW 与 SG 已成为单词表示学习的代表模型. 自提出以来，产生了大量的扩展与应用. 相关衍生工作，将在下一节主要挑战与现有工作中作详细介绍.

3.5　矩阵分解模型

前文所述方法都是基于神经网络的模型. 实际上，矩阵分解同样是得到低维向量表达的重要途径. 本小节主要介绍基于矩阵分解方式学习单词表达的工作.

在基于神经网络的单词表示学习流行之前，最经典成功的单词表示学习模型，当属隐式语义分析(Latent Semantic Analysis/Indexing, LSA/LSI)[57]. LSA 模型，将奇异值分解(Singular Value Decomposition，SVD)应用于单词与文档共现矩阵 $\boldsymbol{X}\in R^{|V|\times n}$，并只保留最大的 k 个奇异值，如下

$$\boldsymbol{X}\approx\boldsymbol{W}\boldsymbol{\Sigma}_k\boldsymbol{D}^{\mathrm{T}}$$

一般使用 $\boldsymbol{W}\boldsymbol{\Sigma}_k$ 作为单词的向量表示. 对于 SVD 分解单词与上下文矩阵，Levy 等人[58]发现 $\boldsymbol{W}(\boldsymbol{\Sigma}_k)^{1/2}$ 在语义相关任务上效果更佳. 而 Caron[59]则建议使用 $\boldsymbol{W}(\boldsymbol{\Sigma}_k)^a$ 形式，其中 a 对结果具有显著影响，需要认真调校. 值得注意的是，Hu 等人[60]发现去掉 LSA 得到表达的第一维后结果会提升，这是因为 LSA 得到的向量的第一维显著大于其它维度.

随后，Huffman 等人[61]概率化 LSI，提出 PLSA (Probabilistic Latent Semantic Indexing)模型，而 Blei 等人[62]将 PLSA 贝叶斯扩展为 LDA(Latent Dirichlet Allocation). 虽然这些模型更多地使用在信息检索场景，关注于文档表达建模，但转变角度，单词同样可以看作关于话题的向量.

除 SVD，典型相关分析(Canonical Correlation Analysis, CCA[63-64])同样被广泛使用于学习单词表示[65-68].

不失一般性地，假设存在随机变量 $\boldsymbol{X}\in R^n$ 和 $\boldsymbol{Y}\in R^m$，CCA 便是寻找映射 $\boldsymbol{a},\boldsymbol{b}$ 使得变换后的 $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X},\boldsymbol{b}^{\mathrm{T}}\boldsymbol{Y}$ 之间的 Pearson 相关系数最大化，形式化表示为

$$(\boldsymbol{a},\boldsymbol{b})=\underset{\boldsymbol{a}\in R^{n\times k},\boldsymbol{b}\in R^{m\times k}}{\arg\max}\frac{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\Sigma}_{XY}\boldsymbol{b}}{\sqrt{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{\Sigma}_{XX}\boldsymbol{a}}\ \sqrt{\boldsymbol{b}^{\mathrm{T}}\boldsymbol{\Sigma}_{YY}\boldsymbol{b}}}$$

其中 $\boldsymbol{\Sigma}$ 代表对应的协方差矩阵.

CCA 的基本思想是优化两个向量，使得它们最大相关化. 使用 CCA 学习单词表示，自然的思路便是用 $\boldsymbol{X}$ 表示单词，而 $\boldsymbol{Y}$ 表示与其关联的上下文，然后便可应用 CCA 将单词与上下文映射到两者最相关的空间中去. 而映射后的向量则可视为单词的新的向量表示.

此外，Lebret 与 Collobert[69]使用 Hellinger 距离作为主成分分析(Principal Component Analysis，PCA)分解单词共现矩阵的损失函数，提出 HPCA (Hellinger PCA)模型. 而受 Mikolov 等人的工作启发，Pennington 等人[21]提出 GloVe 模型，带权重分解单词与其上下文共现的对数矩阵. Shazeer 等人[70]则针对点互信息(Pointwise Mutual Infoermation，PMI)矩阵中缺失值特殊建模，提出了 Swivel 模型学习单词表达，在词频较少的单词表达上取得显著提升.

3.6　模型联系

回顾已有单词分布式表示学习模型，从神经网络到话题模型，一个很自然的问题是：这些模型之间有什么样的联系？它们之间有何相同之处，又有何不同之处？它们在不同的任务上性能比较又如何？本小节主要分析比较现有模型间的联系.

3.6.1　横向组合与纵向聚合

容易发现，所有这些模型都在利用某种上下文

统计信息来学习单词表达.它们都基于同样的一个假设,分布语义假设(Distributional Hypothesis)[71-72],其假设单词的语义来自其上下文.这个假设作为自然语言处理中最重要的假设,同样引起了认知等领域研究人员广泛关注,并得到实证证实[73].

不同的是,一些模型使用文档作为单词的上下文,而另一些则使用单词周边的单词作为上下文.Sun 等人[74]表明,上下文的不同,使得不同模型建模了单词间的不同关系:横向组合关系(Syntagmatic)与纵向聚合(Paradigmatic)关系[75].

关于横向组合关系与纵向聚合关系的示例,如图 4 所示.横向聚合关系,顾名思义是一种横向的关系,指的是两个单词同时出现在一段文本区域中.如图 4 中,"爱因斯坦"与"物理学家"两个词同时出现在一句话中,这两个词间存在着横向组合关系.此关系强调两个词可以进行组合,在句子中往往起到不同的语法作用.而纵向聚合关系,指的是纵向的可替换的关系,如图 4 中的"爱因斯坦"与"费曼".如果两个词在一句话中互换后,不影响句子的语法正确性以及语义合理性,则这两个词间存在纵向聚合关系.纵向聚合关系在形式上表现为,这两个单词出现在相似的上下文环境中,即使这两个单词可能从未共现.

图 4 横向组合与纵向聚合示例

使用文档作为上下文的模型,隐含的假设是,如果两个单词经常同时出现在同一个文档,则这两个单词语义相似.这类模型建模了单词间的横向组合关系,其假设单词与和它共现的单词相似.LSI 以及 LDA 等通常使用在信息检索场景下的模型,都是建模的这类关系.这类模型更多的侧重于单词的话题信息,因而针对如文本分类这类侧重话题的任务,要好于使用纵向聚合关系的模型.如 Tang 等人[76]实验证实,在文本分类任务上,PV-DBOW(Distributed Bag of Words version of Paragraph Vector)要明显优于 SG 模型,其中 PV-DBOW 与 LSI 类似,建模的是单词间的横向组合关系.

而另一类模型,使用单词周边单词作为上下文.其假设,如果两个单词周围的单词相似,则这两个单词语义相似,即使这两个单词可能从未同时出现在一段文本区域中.这类模型建模了单词间的纵向关系,包括 NPLM、LBL、CBOW、SG、GloVe 等.这也是自然语言处理中最常用关系,同时也是分布语义假设最主流的解释.这类模型更加擅长有关单词自身的各项应用.

3.6.2 神经网络与矩阵分解

上文是从模型所利用的基本假设,来分析它们之间的关联.此外,还有一些工作从模型定义的目标函数出发,建立现有模型间的联系.

Levy 和 Goldberg[77]分析表明,SG 模型在使用负采样(SGNS)进行学习的情况下,相当于隐式地在分解单词与上下文之间偏移的点间互信息(Shifted Pointwise Mutual Information, shifted PMI)矩阵.对于 PMI 矩阵,它是自然语言处理领域表示单词语义的一个常用选择[78].随后,Li 等人[79]在表示学习的框架下证明了 SGNS 等同于矩阵分解.

此外,Shi 和 Liu[80]以及 Shazeer 等人[70]的工作都表明,GloVe 模型实际上与使用负采样的 SG(SGNS)模型非常相似,其区别只在于模型中的偏移项以及单词权重的选择.而 Suzuki 和 Nagata[81]则提出了一个统一的框架囊括了这两个模型.

然而,目前这方面的工作主要还是集中于 SG 这样的简单的模型,而对于其他真正意义上的神经网络模型,如 NPLM,我们还缺乏更深入的洞见.

3.7 模型实验比较

上文主要介绍了单词分布式表示学习的发展脉络,并讨论了现有模型之间的联系.然而,面对现有众多模型,一个现实的问题是:在实际使用时,我们应该使用哪种模型?

关于单词分布式表示模型的实验对比,早期始于 Turian 等人[82]的工作.Turian 等人使用了一个 6300 万词的语料学习 HLBL 模型与 C&W 模型的单词表示,并将它们作为额外特征应用于命名实体识别(ner)任务和短语识别(chunking)任务,以比较不同模型对系统性能的提升.之后,Baroni 等人[83]比较了"计数"模型与"预测"模型在若干语义相似度任务中的表现.Baroni 等人将基于统计"词-上下文"共现矩阵,以及在其基础上进行矩阵分解的方法(包括 SVD 与 NMF),统称为"计数"模型;并将基于神经网络的词向量模型(CBOW 与 C&W)统称为"预测"模型.在 28 亿单词的语料上的实验表明,预测模型对在各项指标中比计数模型有显著的优势.同年 Milajevs 等人[84]尝试使用向量按位相加、按位相乘等基本的语义组合方式表示短语以及句子的语

义.然而,实验结果却表明,基于共现矩阵的词表示方法相比神经网络的词向量模型,具有更强的语义组合能力.Levy 等人[58]在中尝试了多种不同的模型参数,发现大部分参数设置的技巧,对基于共现矩阵的模型以及基于神经网络的模型同时有效,并且神经网络模型相比共现矩阵模型在单词相似度估计方面并没有明显优势.

然而,由于前文所述模型都是无监督学习模型,并非针对某些固定任务而设计,因此在各自论文中往往使用不同的语料,不同的任务以及实验设置进行评价.而许多工作又没有公开源码与学习得到的单词表示,因此也难以在一个统一的语料,以相同的设置进行公平比较.因此,现有关于单词表示模型比较的文献,也只是局限于部分模型部分任务.下文尝试从单词相似度、类比任务以及将单词表示作为特征用于外部任务 3 个角度来对比现有代表性模型,分析不同模型的特点,以及其适用场景.

3.7.1　单词相似度

衡量单词的语义相关性,一直是评价单词表示质量的经典任务.其中使用最广泛的数据集当属 WordSim 353(WS 353)数据集[85].该数据集包含了 353 个单词对,其中每一个词对有 13 或者 16 位标注者对其进行 0～10 之间的打分,分数越高表示标注人员认为这两个词语义更相关或者更相似.最终,对于每个词对都可以得到所有标注者的一个平均打分,得到的数据形式如:“tiger cat 7.35”,代表词对“tiger cat”的平均打分为 7.35.对于不同的单词分布式表示模型,在学习出单词的向量表示后,可以通过余弦或欧式距离估计单词对的相似度.此任务的评价标准为,计算标注者对于单词对打分与模型习得表示得到的打分之间的 Spearman 排序相关系数

$$r = \rho_{x,y} = \frac{cov(\boldsymbol{x},\boldsymbol{y})}{\sigma_x \sigma_y}$$

其中:$cov(\boldsymbol{x},\boldsymbol{y})$表示排序列表 $\boldsymbol{x},\boldsymbol{y}$ 之间的协方差;σ_x 与 σ_y 代表了对应的标准差.模型得到的打分与人工标注的打分排序越一致,得分就越高.

除 WS-353 数据集外,本文还选取了另外两个常用数据集评价单词表示.其中 Rare Word(RW)数据集[54]侧重于评价模型学习稀缺单词表示的能力,其包含了2034个单词对.相比其它数据集,RW 包含了更多的词形复杂少见的单词.另一个SimLex-999(SL-999)数据[86]集则修正了 WS-353 混合相关与相似的缺点,专注与单词之间的相似性.相比 WS-353,这个数据集对各个单词表示模型也更难.

表 1　单词相似度实验结果(部分结果来自已有文献[58,70])

模型	语料	维度	WS-353	RW	SL-999
C&W	Wikipedia 2007	50	40.75	38.54	26.99
HPCA	Wikipedia 2012	50	40.26	18.58	15.46
GloVe	Wikipedia 2010	50	55.48	30.57	25.28
CBOW	Wikipedia 2010	50	64.37	40.65	30.87
SG	Wikipedia 2010	50	65.30	39.57	27.66
GloVe	Wikipedia 2010	300	59.18	34.13	32.35
CBOW	Wikipedia 2010	300	67.21	45.19	38.82
SG	Wikipedia 2010	300	70.74	45.55	36.07
SVD	Wikipedia 2013	500	72.15	50.80	42.50
Swivel	Wikipedia 2015+Gigaword5	300	68.20	48.30	40.30

实验结果如表 1 所示.从表中可以看出,CBOW、SG 以及 GloVe 都是非常强有力的模型.相比这些专门用来学习单词表示的模型,C&W 这样用来学习语言模型以及监督任务的模型得到的单词表示并不能很好的捕捉单词见无监督的语义相似信息.此外 C&W 模型训练速度也远比 CBOW 等模型要慢,在 Wikipedia 2007 的语料上,Collobert 和 Weston 训练了两个月才得到最终的单词表示.对比 SVD、GloVe 这样的矩阵分解模型与 CBOW、SG 这些神经网络模型,可以发现在此任务上两者之间并没有很显著的差异.但是两者的适用场景很不一样.如 Jeffrey 等人[21]所述,矩阵分解速度要快于 CBOW 以及 SG 模型.但是在内存占用上矩阵分解模型要远高于 CBOW 以及 SG 模型.另外,CBOW 和 SG 使用的还是一种在线的训练方式,并且可以方便地利用已有单词表示作为初始值重新更新单词表示.

此外,对比在 RW 数据集与其他数据集上的表现可以发现,CBOW、SG 以及 Swivel 模型在稀缺词上的表现都远好于 GloVe 模型.这方面更详细的比较可见 Shazeer 等人[70]的工作.这是因为 GloVe 相比其他模型,并没有使用负采样技术.在学习过程中,模型并不会对那些没有共现却具有相似表示的单词进行惩罚.而 CBOW 等模型因为考虑了数据中并没有出现的共现信息,因此可以在词频较少的单词上具有较好的结果.

3.7.2　单词类比

除单词相似度任务外,单词的类比任务同样被广泛用于评价单词表示质量.单词类比任务最早由 Mikolov 等人[87]提出以量化评价单词对关系的相似性,主要数据集由 Mikolov 等人[4]在工作中提出.该数据集包含了大量的类似于“a 之于 b 正如 c 之于 __”这样的问题,其中缺失的单词需要寻找整个词汇表中来回答.回答这样的问题,需要找到一个单词的向量 $vec(x)$,它尽可能的接近 $vec(b)-vec(a)+$

$vec(c)$：

$$\arg\max_{\substack{x\in V,x\neq a\\ x\neq b,x\neq c}}(vec(b)-vec(a)+vec(c))vec(x)$$

只有当单词 x 正好是缺失的单词时，此问题才被判为回答正确. 最终的评价标准则是回答正确的问题占所有问题的百分比.

此数据集包含了 5 类语义类比任务以及 9 类句法类比任务. 其中，语义类比任务包含了 8869 个问题，大体上都是关于人物和地点的问题，如“Beijing：China～Paris：France”. 而句法类任务则包含了 10675 个关于词形变换类的问题，如“good：better～bad：worse”.

实验结果如表 2 所示. 可以发现，相比 SVD 以及 HPCA 模型，CBOW 和 SG 在单词类比任务上具有明显优势. 这也印证了 Arora 等人[88]从数学上对于 SG 等模型捕捉单词间的类比关系的解释. SG 这类线性模型相比 C&W 等非线性模型更能捕捉单词间的线性关系. 结合上文单词相似度实验结果，可以发现 Word2Vec 的两个模型，尤其是 CBOW 相比 GloVe 具有明显优势. 虽然 Shi 和 Liu[80]以及 Shazeer 等人[70]的工作都表明，GloVe 模型与使用负采样的 SG(SGNS)模型非常相似，其区别只在于模型中的偏移项以及单词权重的选择. SGNS 相当于 GloVe 使用单词词频的对数作为便宜项，这是一个非常好的设置，而 GloVe 在实际优化时，未必能学到这么好的参数设置，因为在实际使用时未必能达到 SG 以及 CBOW 模型的性能.

表 2 单词类比实验结果(部分结果来自已有文献[58,70])

模型	语料	维度	Sem	Syn	Total
C&W	Wikipedia 2007	50	40.75	38.54	26.99
HPCA	Wikipedia 2012	50	3.36	10.42	7.20
GloVe	Wikipedia 2010	50	56.60	43.53	49.46
CBOW	Wikipedia 2010	50	60.86	50.55	55.23
SG	Wikipedia 2010	50	50.27	43.93	46.81
GloVe	Wikipedia 2010	300	79.85	61.15	69.64
CBOW	Wikipedia 2010	300	79.65	68.54	73.58
SG	Wikipedia 2010	300	77.16	65.31	70.69
SVD	Wikipedia 2013	500	—	—	55.40
Swivel	Wikipedia 2015+Gigaword5	300	—	—	73.90

3.7.3 单词表示用作特征

除上述两种利用单词自身属性评价单词表示的方法外，单词表示往往也被放到外部任务中进行评价. 本文选用两个有代表性的任务：(1)将单词向量表示作为现有系统的额外特征，完成名词短语识别任务；(2)将单词向量表示作为唯一特征，应用于情感分类任务.

对于名词短语识别任务，利用学习得到的单词分布式表示作为辅助特征，应用于 CRF 模型. 所有模型均在 CoNLL-2000 公开任务[89]上使用 F_1 指标进行评价. 对于情感分类任务，使用 Mass 等人[6]公开的数据集. 此数据集包含了均衡的正负两类样本，共计50000条电影评论信息. 该实验直接使用文本中各词词向量的加权平均值作为文档的表示，以此作特征，使用 Logistic 回归[90]学习分类模型，并使用 10 重交叉验证进行评价，评价标准使用 F_1.

实验结果如表 3 所示，其中 TSCCA 代表了 EigenWord 中的两步 CCA 模型[66-67]. 可以发现在句法相关任务上 C&W 模型要优于其他模型. 相比其他模型忽略了单词间的顺序信息，C&W 建模了单词见的顺序信息. Landauer[92]分析文本中约有 20%的语义来自于词序，而剩下部分来自词的选择. Ling 等人[93]在 CBOW 以及 SG 模型加入顺序信息，实验表明在依存解析以及词形标准任务上都取得了性能提升. 另一方面，在语义相关任务上，可以发现 CBOW 要显著好于其他模型. 整体而言，可以发现不同模型在这些外部任务上的性能差距都很小，这是因为这些模型都没有争对特定的任务进行设计. 此外，可以发现单词表示学习模型在单词自身属性评价结果与外部具体任务上的性能并不一致，如 C&W 在名词短语识别上好于 CBOW，但是在单词相似度与类比实验上，都远差于 CBOW. 这些都对单词表示模型的评价提出了挑战，这方面的具体讨论可见 4.5 节.

表 3 单词表示作为特征实验结果(结果来自已有文献[91])

模型	名词短语识别	情感分类
C&W	94.53	72.37
HPCA	94.48	69.45
TSCCA	94.53	75.02
GloVe	94.28	74.87
CBOW	94.32	75.78

3.8 小 结

单词分布式表示在取得巨大成功的同时，也伴随着各种争议. 如 Turian 等人[82]表示现有单词分布式表示在实际应用中所取得的提升，并没有 Brown 聚类[43]作为单词的表示所取得的提升大①. Levy 等人[58]则表示 SGNS 以及 GloVe 模型实际上并没有比传统的 PMI 矩阵以及对其做 SVD 分解得

① Brown 聚类通过极大化临近单词类别间的互信息，层次地聚类单词形成二叉树，树的叶子节点是单词，而中间节点则是类别. 而这些聚类也可以被用作单词表示[82,94,149-150].

到的低维表示在单词相似度任务上更优秀，更多的只是超参的调优选择.

而另一方面，也有许多工作认为以 SG 为代表的这些分布式表示模型要优于传统的简单的基于共现统计的表示. 如 Baroni 等人[83]系统性地比较了基于上下文 PMI 矩阵的单词表示与 CBOW 模型，结果显示 CBOW 几乎在所有任务上超越 PMI 矩阵表示. 另外，Guo 等人[94]则反驳了 Turian 等人[82]的观点，显示分布式表示比 Brown 聚类具有更强的表示能力.

从上可见，当前对于分布式表示的性能依然存在着许多争论，并无定论，还有待进一步深入研究.

4　主要挑战以及已有解决方案

以 Word2Vec 为代表的单词表示学习模型，已经在单词相似度建模、类比等任务取得了瞩目的成果. 但是，单词的表示学习依然面临着诸多挑战. 此节，旨在总结目前单词表达面临的主要挑战，以及相应的已有工作.

4.1　多义词表示

前文所述工作，都是将单词表示为一个向量. 然而，单词中存在着大量的一词多义现象. 简单地将所有不同语义编码在一个向量中，则会给后续应用带来诸多问题. 如由于距离的三角不等式的存在①，两个自身语义不相似，但是都与另外一个多义词不同语义相似的单词，在表示空间中会被不恰当的拉的更近.

对此问题，最简单直接的方式便是使用多个表达表示多义词的每个语义. 早在文献[48,50]中，研究人员已经开始关注到单词表示中的多义词问题，在输出层对多义词设置多个编码，但是单词的多义性并没能在最终的单词表示中得以体现. Huang 等人[53]使用预先学好的单词上下文的向量表达进行 k-means 聚类，依据聚类结果对训练语料中单词重新标注不同词义，从而学出多义词的不同语义表示. 然而，Huang 等人的模型基于 C&W 模型的学习框架，其复杂度依然较高，耗时一周才在十亿单词的语料上学习出 30 000 词汇中 6000 个单词的多向量表示.

而随后的一系列工作，大体上都是基于 Word2Vec 进行扩展，学习多义词的多向量表示. Tian 等人[95]在 SG 模型的基础之上，使用概率混合模型，学习单词的多向量表示. 前面所述工作，都假设所有单词具有同样的语义个数，这显然并不合理. 为解决此问题，Neelakantan 等人[96]在 SG 模型的基础上，提出一种非参模型 NP-MSSG，对多义词自动学出不同个数的向量表示. 而 Chen 等人[97]则是利用外部资源，根据 WordNet[98]定义好的单词的多义列表，学习多义词的多向量表达.

其他相关工作，还包括：Qiu 等人[99]针对单词的每个不同词性分别学习不同的表达；Liu 等人[100]则在 SG 模型的基础上引入 LDA 的话题信息，学习话题相关的单词表达；相比文献[100]，Liu 等人[101]使用张量神经网络扩展 SG，提出一个更加通用的模型 NTSG(Neural Tensor Skip-Gram)学习话题相关的单词表示.

而对于多义词表示最重要的问题是，将多义词用多个向量表示，真的能提升实际 NLP 应用的效果么？上述工作并没有正面回答此问题. 为回答此问题，Li 和 Jurafsky[102]测试了主流多义词表示模型(如文献[53,96])，发现在词性标注和语义关联任务上可以提升效果. 而对于语义关联任务，当使用如长短项记忆模型(Long-Short Term Memory，LSTM)这类复杂模型时，这种提升也被模型能力抹平. 至于命名实体识别以及情感分析任务，多向量表示并没能提升性能.

因此，多义词的表示学习与应用，依然是一个有待解决的挑战.

4.2　稀缺单词表示

现有的单词表示学习模型，大体都是利用大量纯文本数据，根据单词的上下文分布来学习单词表示. 这种模式的一个挑战在于，对于那些出现次数较少的单词，其周围的上下文往往不足以学习出一个好的表示. 针对此问题，已有一些工作，利用单词自身结构所携带信息，帮助学习那些稀缺单词的表示.

Luong 等人[54]在 C&W 的框架下，使用递归神经网络(Recursive Neural Network，RecursiveNN)建模单词内部结构，以更好地学习那些出现次数较少的词形复杂的单词表达. Botha 和 Blunsom[103]将词素信息引入 LBL 模型帮助学习单词表达和语言模型. Qiu 等人[104]扩展 CBOW 模型，使用上下文以及其词素的表达预测目标单词及其词素来学习单词以及词素的表达. 而后，Sun 等人[105]扩展 CBOW 模型，在使用上下文预测目标词的同时还使用目标词词素预测它. 相似的思路，被对称地用来扩展 SG 模

① 对于距离度量 d，有 $d(a,c)\leq d(a,b)+d(b,c)$.

型.相比文献[104]与[105]显式地建模了单词与自身词素之间的交互.

上述已有工作,都是基于单词的词素信息,试图使用共享词素的单词表示来帮助那些稀缺词的表示学习.然而对于那些出现次数少,而又不是由词素构成的单词,利用词素办法也难以作为.针对此问题,已有工作尝试利用字符级信息,增强单词学习表示[106].这依然是一个富有挑战的研究方向.

4.3 细粒度语义建模

现有基于分布语义假设,仅利用上下文统计信息的模型,难以学习单词的细粒度语义.如反义词通常可以出现在相似的上下文中,而现有模型大体难以区分这些单词的向量表示.再有 Rubinstein 等人[107]实验发现,使用分布语义假设学习得到的单词表达,不能很好的捕捉单词的属性信息.而现实中,已经存在许多先验知识库描绘了单词间的关联信息,如 WordNet[98]和 FrameNet[108].近年来,已有一些工作尝试利用这些先验知识帮助学习单词表达.

第1类工作,通过修改已有模型的目标函数,引入知识库信息(如单词间关系、单词的属性)作为约束,优化单词表达的学习.如使用 WordNet 中同义词关系,使得这些同义单词的向量表达更相似.Yu 和 Dredze[109]在 CBOW 的基础上,提出 RCM(Relation Constrained Model)使用单词去预测知识库中相关的单词的表达.但此工作忽略了知识库中所标注的单词间关系的不同.Bian 等人[110]尝试将 Longman 词典①、WordNet 等知识库中词法、句法以及语义等外部知识作为附加输入或监督信息引入 CBOW 模型.而后,Xu 等人[111]利用知识库中单词关系和属性信息作为距离约束引入 SG 模型.Liu 等人[112]将外部知识作为不等式约束来扩展 SG 模型.

而另一类工作则尝试利用知识库,直接改善已有单词表达.Faruqui 等人[113]使用单词间关联信息构建单词间网络,并在图上使用信念传播(Belief Propagation,BP)更新单词的表示,使得关联单词的向量表示更接近.

4.4 单词表示的解释性

虽然单词的分布式表达在近些年取得了令人瞩目的成功,但它的可解释性一直是被诟病的严重问题.对于这种稠密的实数向量表示,我们并不知道它的每一维代表什么,或是多个维度一起代表什么,也不知道这些维度上的值的意义.如我们并不知道"狗"的向量表示的哪些维度代表了它的性别,也不知道什么样的值代表"雌"或"雄".然而,即使我们知道哪些维度表征性别,它依然是难以解释的,因为这些维度在所有的单词上都会有非零值,包括那些无性别的事物.这种表示方式,既是难以解释,也是不经济的[114-115].

以往工作显示,稀疏化是提高表示可解释性的一个可行方案.关于人的视觉研究发现,初级视觉皮质(primary visual cortex,V1)中的神经元具有一种分布式的稀疏表示[116-117].而语言方面的研究也发现,人们倾向于只使用 20～30 个特征去描述一个单词[118].

近年来,有一些工作尝试使用稀疏向量来提高单词表示的可解释性.

Murphy 等人[119]使用非负稀疏编码以提高矩阵分解得到的单词表示的可解释性,并与 LSA 做了翔实对比.Faruqui 等人[120]则直接对 Word2Vec 学到的单词表示做稀疏编码(Sparse Coding,SC),结果显示,相比原始表达,分解后的稀疏表示在提升可解释性的同时,在实际外部应用(如文本分类)上性能也得到了提升.与文献[120]后处理的方式不同,Sun 等人[121]直接在 CBOW 模型中引入稀疏约束,直接学习稀疏表示.

此外,Luo 等人[122]则受非负矩阵分解启发,在 SG 模型中引入非负约束,结果显示同样可以提升表达的可解释性.

4.5 单词表示的评价

单词表示的评价,同样是单词表示学习领域面临的一个挑战.常见的评价方式可分为内源(intrinsic)评价与外源(extrinsic)评价.

内源评价,主要有单词相似度以及类比任务.相似度任务,主要衡量模型学出的单词表示编码的单词间相似度与人给出的单词相似度之间的相关性.除去普通的单词相似度测试集,如 WordSim353[85]、SimLex 999[86]外,还有针对特殊场景设计的测试集,如评价多义词表示的 SCWS(Stanford Contextual Word Similarity)[53],评价稀缺词相似度的 Rare Word[54].

单词类比任务来源于 Mikolov 等人[87]发现使用循环神经网络(Recurrent Neural Network,RNN)学到的单词表示不但可以编码单词间的相似度,还可以编码单词对间的相似度,也就是 Mikolov 所谓的语言正则性(linguistic regularities),又或者 Turney 的关系相似度(relational similarities)[123].如"北京之于中国,相当于巴黎之于法国".Mikolov

① http://www.longmandictionariesonline.com.

发现，使用简单的向量运算，便可以回答语法、语义上的这类类比问题. 如上面的示例可通过向量减法捕捉：vec(北京)－vec(中国)＝vec(巴黎)－vec(法国)，这里 $vec(x)$代表单词 x 的向量表示. 如图 5 所示，3 个首都与国家对之间的向量相减后得到了近似平行的向量. 这表明，这种向量表示捕捉到了单词之间的语义关联. 而后，单词的类比任务被广泛用于评价单词表示的好坏.

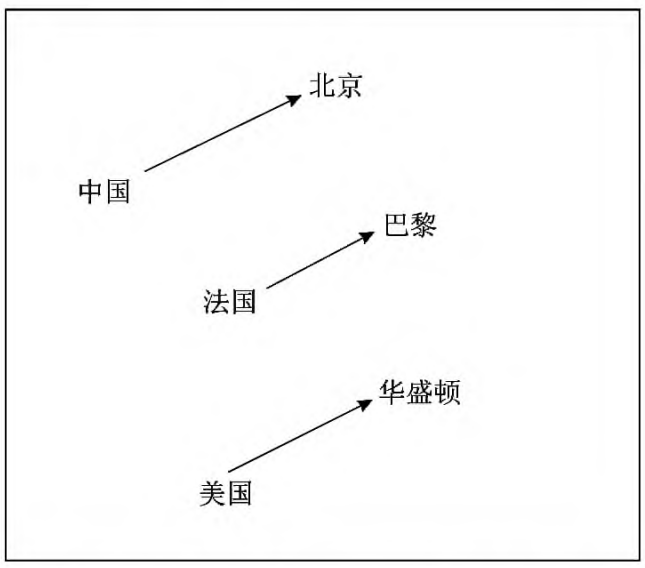

图 5　单词类比示意图(利用 PCA 将单词向量表示降维到 2 维后示意图，途中箭头所示为使用对应的首都的向量减去国家向量表示后得到的向量)

上述两种方式，都仅仅是评价单词表示的内在属性. 而 Tsvetkov 等人[124]发现，在外源任务上的好坏，并不能反映单词表示在外部实际应用上的好坏. 学习单词表示，最终目的毕竟还是为了实际应用. 因此也使用命名实体识别、单词极性分类等监督任务[82,125]来评价单词表示的好坏.

然而，一则若以监督任务的结果为指标，那么修改任务模型本身所能带来的提升往往远高于不同单词表示之间的差异. 二则很难说明监督任务所使用的模型对于单词表示有无偏好. 如 Schnabel 等人[91]发现不同的任务倾向于不同模型得到的单词表示，外部监督任务并不适合作为评价单词表示质量的桥梁.

因此，合理评价单词表示的质量，依然是一个挑战. 一个初步的进展是，Tsvetkov 等人[124]使用与手工构建的单词语言学特征表示[126]的相关系数作为评价单词表示质量的指标. 他们同时实验证实了这种指标比单词相似度任务指标，与实际应用中的性能具有更高的相关性.

5　单词表示未来研究方向展望

近年来单词的表示学习，已引起研究人员的广泛关注，并在很多任务中展现了巨大的应用潜力. 对于其面临的挑战，也已经提出了许多探索工作、解决方案. 本节将对单词表示学习的未来进行展望.

5.1　面向具体应用的单词表示

无监督的单词表示，并没有考虑具体任务信息，因此难以发挥单词表示的最大效用，也造成单词表示评价的困难. 未来，结合具体应用特点，在单词表示学习中引入任务相关的信息，将是一个研究热点.

在学习过程中，考虑任务特点，主要可分为两类方式.

(1) 直接在单词学习目标函数中引入任务信息. 如为更好应用于情感分类任务，在学习单词表示时，引入情感极性信息，学习的到带有情感极性的单词表示[5-6,127]. 此外，Tang 等人[76]在学习单词表达时，引入标签信息，以增强单词表示在分类任务中的判别能力.

(2) 针对任务特点，设计特定的模型. 如 Ling 等人[93]针对句法类任务，将序列信息引入 CBOW 和 SG 模型. Chen 等人[128]以及 Li 等人[129]根据中文的特点，改进 CBOW 与 SG 模型，以更好的学习中文字、词的表示.

5.2　单词分布式表示应用的研究

自 Word2Vec、GloVe 等高效的单词表示学习算法提出，关于单词表示学习算法本身的研究已经放缓. 如近两年关于单词表示学习的算法，大体上都是基于 Word2Vec 的扩展. 这主要是因为现有算法已经可以高效处理大规模数据，支撑实际应用. 现有使用单词表示的工作，大多集中于简单地将单词的向量表示作为系统的输入. 而未来的研究热点，将主要集中于如何利用单词分布式表示的特点，更好地解决实际应用问题.

词袋模型是现有信息检索、话题模型等众多研究的基础假设，而其本身则以独热表示为基础. 相比于独热表示，单词的分布式表示编码了单词语义信息先验知识. 因此，如何利用分布表示，改进这些模型，甚至提出基于分布式表示的框架，将是一个有意义的研究方向.

如近期关于在话题模型中引入单词分布式表示的工作便是这方面探索的典型. 包括，Nguyen 等人[13]利用单词表示修改 LDA 与 DMM(Direchlet Multinomial Mixture)模型中话题到单词之间的映射选择过程. Das 等人[14]则让 LDA 产生单词的向量表示而非单词标示，以克服 LDA 原本词袋假设带来的缺陷. 其他工作还包括，利用单词表示在话题

模型中引入序列信息[15]，使用单词表示改进话题学习时抽样过程[130].

对于信息检索系统，其中一个关键的问题在于因独热表示而引起的词汇不匹配问题(vocabulary mismatch). 它是指文档虽然相似但使用的是不同词汇或查询使用的词汇与相关文档使用词汇相似却不一致而导致的丢失相关结果问题. 单词的分布式表示，则打破了独热表示的约束，非常适合解决此问题. Ganguly 等人[131]、Nalisnick[19] 等人以及 Vulic 等人[132] 的工作，在此方向做出了初步的探索. 此外，Mitra 等人[133] 与 Grbovic 等人[134] 的工作，则探索将单词分布式表示应用于查询扩展改写.

此外，对于单词分布式表示编码类比关系的应用，也是一个值得探索的方向. Fu 等人[135] 利用单词分布式表示所展现的单词间类比关系学习单词间的层次结构. Soricut 与 Och[136] 则利用此特性学习归纳单词的词形规则.

5.3 多源信息融合的单词表示

针对第 4 节所述诸多挑战，都可以融合多源信息来尝试解决，如表达解释性、稀缺单词表示、单词细粒度语义.

如对于单词的分布式表示缺乏解释性，Fyshe 等人[137] 利用功能性磁共振成像(Functional Magnetic Resonance Imaging, fMRI) 和脑磁图(Magnetoencephalography) 数据，非负稀疏联合分解文本与大脑数据，以提高表示的解释性. Faruqui 和 Dyer[126] 使用多个单词知识库构建了一个可解释的特征表示. 而将现有各种单词表示，与此可解释的表示进行对齐，或许可以解释每一个特征所对应的分布式表示的模式. 但文献[126]所得到的最终表示中冗余噪声较多且高维稀疏，如何对齐这两种表达，亦是一个问题.

此外，多源信息，不仅仅限于前文所提及的词根信息、知识库，还可以是跨语言文本数据. 如 Faruqui 和 Dyer[138] 使用 CCA 学习平行语料单词表示，显著提升英文单词表示质量.

目前，对于可利用的资源以及如何利用都只是处于初步阶段，将来必定还会有更多工作涌现.

5.4 单词分布式表示属性研究

虽然研究人员已经提出了大量的单词表示学习算法，并将这些向量表示广泛应用于自然语言处理的各种应用. 但是对于这些表示具有什么样的属性，编码了什么样的信息，以及如何应用这些特性，我们还知之甚少.

从 Mikolov 等人发现对单词分布式表示做简单的向量运算可以揭示单词间的类比关系，到 Schnabel、Wilson、Schakel 等人[91,139-140] 发现现有单词表示学习算法得到的向量，或多或少都编码了单词的词频信息. 再到 Arora 等人[88] 尝试从数学上解释为什么以 Word2Vec 为代表的单词表示学习算法可以捕捉单词间的类比关系. 研究人员在一步步深入探索单词分布式表示的特性.

正如类比性质的发现，启发后续工作[135] 学习单词间的层次结构以及文献[136]利用学习归纳单词的词形规则. 对于单词表示自身性质更好的理解，可以促进相应的应用研究. 而对于单词分布式表示自身的研究还处于刚刚起步的阶段，这也是一个未来值得深入研究的领域.

5.5 其他研究方向

除以上几个研究方向外，还有很多关于单词表示学习的研究工作亟待展开.

(1) 单词表示的组合. 如何利用单词表示，组合得到短语、句子以及文档的表达，也是未来一个研究热点. 目前，此方向主要工作集中于如何通过单词组合得到短语的表示. 如 Socher 等人[8,28,141-142] 使用递归神经网络建模单词组合成短语的过程. Mitchell 以及 Blacoe 等人[143-144] 对单词向量表示的组合方式(如加法、乘法、张量乘法以及卷积等)进行了全面的实验对比，显示简单的加法与乘法就可以表现的不错. 然而 Samuel 与 Tenenbaum[145] 最近则表示，无论是加法、乘法，还是基于递归神经网络的模型，都不能像人那样捕捉短语的组合模式. 因此，单词表示的组合还是一个众说纷纭的开放问题，将是未来的一个研究热点. 而如文献[145]所言，LSTM 不失为一个值得探索的可能方案.

(2) 多语言单词表示. 现有工作多数都是针对英文文本，而针对其他语言的单词表示目前还处于起步阶段，有待继续探索. 如对于中文字词表示学习的探索[128-129]，还有词形规则丰富的语言的单词表示[103]，常用于翻译领域的双语单词表示[12,146].

(3) 概率化表示. 分布式表示可以看作将单词表示为空间中的一个点，但这种表示并不能自然地表示单词本身含义的不确定性. 此外，基于向量表示的相似度计算往往使用余弦或欧式距离计算单词间的相似度. 这种相似度度量并不能捕捉单词间的不对称关系. 针对此问题，Erk[147] 尝试将单词表示为空间中的区域. 进一步地，Vilnis 和 McCallum[148] 则将单词表示为多维正态分布. 将单词表示为概率分布，还是一个崭新的研究方向. 相比于将单词表示为

低维实数向量，现有工作还处于很初步的阶段，许多问题还有待发掘研究，如表示为概率，是否还能像向量一样，通过简单的操作表示单词间的类比关系. 但概率分布表示相比于向量表示的优点，预示着这可能是未来单词表示范式的方向.

6　结束语

单词表示，作为自然语言处理等领域的一个基本问题，一直是相关领域的研究热点. 而单词的分布式表示，因其相比于独热表示的诸多优点，近年来亦广受关注，并取得众多进展. 本文首先梳理了近年来单词分布式表示代表性方法的发展脉络，并从单词间组合与聚合关系出发，分析现有模型间的联系，将这些模型统一在分布语义假设框架下. 此外，本文还对单词分布式表示面临的主要挑战、已有解决方案以及未来研究方向进行了总结.

期待更多研究者加入到单词表示学习的研究队伍中，也希望本文对于单词表示学习在国内的研究发展提供一些帮助.

参　考　文　献

[1] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828

[2] Manning C D. Computational linguistics and deep learning. Computational Linguistics, 2015, 41(4): 701-707

[3] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model. Journal of Machine Learning Research, 2003, 3: 1137-1155

[4] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space//Proceedings of Workshop of ICLR. Scottsdale, USA, 2013: 1-12

[5] Tang D, Wei F, Yang N, et al. Learning sentiment-specific word embedding for twitter sentiment classification// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, 2014: 1555-1565

[6] Maas A L, Daly R E, Pham P T, et al. Learning word vectors for sentiment analysis//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, USA, 2011: 142-150

[7] Socher R, Manning C D, Ng A Y. Learning continuous phrase representations and syntactic parsing with recursive neural networks//Proceedings of the Deep Learning and Unsupervised Feature Learning Workshop of NIPS 2010. Hyatt RegencyCanada, 2010: 1-9

[8] Socher R, Bauer J, Manning C D, et al. Parsing with compositional vector grammars//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria, 2013: 455-465

[9] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning//Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland, 2008: 160-167

[10] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv: 1409.0473, 2014

[11] Mikolov T, Le Q V, Sutskever I. Exploiting similarities among languages for machine translation. arXiv preprint arXiv: 1309.4168, 2013

[12] Zou W Y, Socher R, Cer D, et al. Bilingual word embeddings for phrase-based machine translation//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 1393-1398

[13] Nguyen D Q, Billingsley R, Du L, et al. Improving topic models with latent feature word representations. Transactions of the Association for Computational Linguistics, 2015, 3: 299-313

[14] Das R, Zaheer M, Dyer C. Gaussian LDA for topic models with word embeddings//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. Beijing, China, 2015: 795-804

[15] Yang M, Cui T, Tu W. Ordering-sensitive and semantic-aware topic modeling//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, USA, 2015: 2353-2359

[16] Manning C D, Raghavan P, Schütze H. Introduction to Information Retrieval. New York, USA: Cambridge University Press, 2008

[17] Croft B, Metzler D, Strohman T. Search Engines: Information Retrieval in Practice. USA: Addison-Wesley Publishing, 2009

[18] Hill F, Cho K, Korhonen A. Learning distributed representations of sentences from unlabelled data//Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics. San Diego, California, 2016: 1367-1377

[19] Nalisnick E, Mitra B, Craswell N, et al. Improving document ranking with dual word embeddings//Proceedings of the 25th International Conference Companion on World Wide Web. Republic and Canton of Geneva, Switzerland, 2016: 83-84

[20] Hinton G E, McClelland J L, Rumelhart D E. Distributed representations//Rumelhart D E, McClell J L, PDP C eds. Research Group. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1. Cambridge, USA: MIT Press, 1986: 77-109

[21] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1532-1543

[22] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch. Journal of Machine Learning Research, 2011, 12: 2493-2537

[23] Mnih A, Hinton G. Three new graphical models for statistical language modelling//Proceedings of the 24th International Conference on Machine Learning. Oregon, USA, 2007: 641-648

[24] Hill F, Cho K, Korhonen A, et al. Learning to understand phrases by embedding the dictionary. Transactions of the Association for Computational Linguistics, 2016, 4: 17-30

[25] Cho K, van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder—Decoder for statistical machine translation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1724-1734

[26] Lebret R, Collobert R. The sum of its parts: Joint learning of word and phrase representations with autoencoders. ICML Deep Learning Workshop. Lille, France, 2015

[27] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of wrds and phrases and their compositionality//Proceedings of the Advances in Neural Information Processing Systems 26. Lake Tahoe, USA, 2013: 3111-3119

[28] Socher R, Huval B, Manning C D, et al. Semantic compositionality through recursive matrix-vector spaces//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA, 2012: 1201-1211

[29] Paccanaro A, Hinton G E. Learning distributed representations of concepts using linear relational embedding. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(2): 232-244

[30] Hu B, Lu Z, Li H, et al. Convolutional neural network architectures for matching natural language sentences//Proceedings of the Advances in Neural Information Processing Systems 27. Montréal, Canada, 2014: 2042-2050

[31] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, 2014: 655-665

[32] Le Q, Mikolov T. Distributed representations of sentences and documents//Proceedings of the 31st International Conference on Machine Learning. 2014: 1188-1196

[33] Tang J, Qu M, Wang M, et al. LINE: Large-scale information network embedding//Proceedings of the 24th International Conference on World Wide Web. Florence, Italy, 2015: 1067-1077

[34] Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online learning of social representations//Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2014: 701-710

[35] Miikkulainen R, Dyer M G. Natural language processing with modular PDP networks and distributed lexicon. Cognitive Science, 1991, 15: 343-399

[36] Rumelhart D E, McClelland J L, C. PDP Research Group. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge, MA, USA: MIT Press, 1986

[37] Bengio Y, Ducharme R, Vincent P. A neural probabilistic language model//Proceedings of the Advances in Neural Information Processing Systems 13. Vancouver, Canada, 2001: 932-938

[38] Xu W, Rudnicky A. Can artificial neural networks learn language models?//Proceedings of the 6th International Conference on Spoken Language Processing. Beijing, China, 2000: 202-205

[39] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504-507

[40] Schwenk H, Gauvain J-L. Connectionist language modeling for large vocabulary continuous speech recognition//Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Orlando, USA, 2002, 1: 765-768

[41] Ponte J M, Croft W B. A language modeling approach to information retrieval//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia, 1998: 275-281

[42] Kneser R, Ney H. Improved backing-off for M-gram language modeling//Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. Michigan, USA, 1995, 1: 181-184

[43] Brown P F, deSouza P V, Mercer R L, et al. Class-based N-gram models of natural language. Computational Linguistics, 1992, 18(4): 467-479

[44] Bengio Y, Sénécal J-S. Quick training of probabilistic neural nets by importance sampling//Proceedings of the Conference on Artificial Intelligence and Statistics. Key West, USA, 2003

[45] Mnih A, Teh Y W. A fast and simple algorithm for training neural probabilistic language models//Proceedings of the 29th International Conference on Machine Learning. Edinburgh, Scotland, 2012: 1751-1758

[46] Mnih A, Kavukcuoglu K. Learning word embeddings efficiently with noise-contrastive estimation//Proceedings of Advances in Neural Information Processing Systems 26. Lake Tahoe, USA, 2013: 2265-2273

[47] Gutmann M U, Hyvärinen A. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. Journal of Machine Learning Research, 2012, 13(1): 307-361

[48] Morin F, Bengio Y. Hierarchical probabilistic neural network language model//Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics. Barbados, 2005: 246-252

[49] Goodman J. Classes for fast maximum entropy training// Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. UT, USA, 2001: 561-564

[50] Mnih A, Hinton G E. A scalable hierarchical distributed language//Proceedings of the Advances in Neural Information Processing Systems 21. Vancouver, Canada, 2008: 1081-1088

[51] Mikolov T, Kopecky J, Burget L, et al. Neural network based language models for highly inflective languages// Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Taipei, China, 2009: 4725-4728

[52] Okanohara D, Tsujii J. A discriminative language model with pseudo-negative samples//Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic, 2007: 73-80

[53] Huang E H, Socher R, Manning C D, et al. Improving word representations via global context and multiple word prototypes //Proceedings of the 50th Annual Meeting of the Association for Computational Linguistic. Stroudsburg, USA, 2012: 873-882

[54] Luong M-T, Socher R, Manning C D. Better word representations with recursive neural networks for morphology// Proceedings of the 17th Conference on Computational Natural Language Learning. Sofia, Bulgaria, 2013: 104-113

[55] Ji S, Yun H, Yanardag P, et al. WordRank: Learning Word Embeddings via Robust Ranking. arXiv preprint arXiv: 1506.02761, 2015

[56] Lazaridou A, Dinu G, Baroni M. Hubness and pollution: delving into cross-space mapping for zero-shot learning// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. Beijing, China, 2015: 270-280

[57] Deerwester S, Dumais S T, Furnas G W, et al. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990, 41(6): 391-407

[58] Levy O, Goldberg Y, Dagan I. Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics, 2015, 3: 211-225

[59] Caron J. Experiments with LSA scoring: Optimal rank and basis//Berry M W. Computational Information Retrieval. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2001: 157-169

[60] Hu X, Cai Z, Franceschetti D, et al. LSA: The first dimension and dimensional weighting//Proceedings of the 25th Annual Conference of the Cognitive Science Society. Boston, USA, 2003: 587-592

[61] Hofmann T. Probabilistic latent semantic indexing// Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berkeley, USA, 1999: 50-57

[62] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. Journal of Machine Learning Research, 2003, 3: 993-1022

[63] Hotelling H. Relations between two sets of variates. Biometrika, 1936, 28(3-4): 321-377

[64] Hotelling H. The most predictable criterion. Journal of Educational Psychology, 1935, 26(2): 139-142

[65] Stratos K, Collins M, Hsu D. Model-based word embeddings from decompositions of count matrices//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. Beijing, China, 2015: 1282-1291

[66] Dhillon P S, Foster D P, Ungar L H. Eigenwords: Spectral word embeddings. Journal of Machine Learning Research, 2015, 16: 3035-3078

[67] Dhillon P, Rodu J, Foster D P, et al. Two step CCA: A new spectral method for estimating vector models of words// Proceedings of the 29th International Conference on Machine Learning. Edinburgh, Scotland, 2012: 1551-1558

[68] Dhillon P, Foster D P, Ungar L H. Multi-view learning of word embeddings via CCA//Proceedings of the Advances in Neural Information Processing Systems 24. Granada, Spain, 2011: 199-207

[69] Lebret R, Collobert R. Word embeddings through hellinger PCA//Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden, 2014: 482-490

[70] Shazeer N, Doherty R, Evans C, et al. Swivel: Improving Embeddings by Noticing What's Missing. arXiv preprint arXiv: 1602.02215, 2016

[71] Firth J R. A synopsis of linguistic theory 1930-55. Studies in Linguistic Analysis (special volume of the Philological Society), 1957, 1952-59: 1-32

[72] Harris Z. Distributional structure. Word, 1954, 10(23): 146-162

[73] Mcdonald S, Ramscar M. Testing the distributional hypothesis: The influence of context on judgements of semantic similarity// Proceedings of the 23rd Annual Conference of the Cognitive Science Society. Edinburgh, Scotland, 2001: 611-616

[74] Sun F, Guo J, Lan Y, et al. Learning word representations by jointly modeling syntagmatic and paradigmatic relations// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. Beijing, China, 2015: 136-145

[75] Sahlgren M. The distributional hypothesis. Italian Journal of Linguistics, 2008, 20(1): 33-54

[76] Tang J, Qu M, Mei Q. PTE: Predictive text embedding through large-scale heterogeneous text networks//Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, Australia, 2015: 1165-1174

[77] Levy O, Goldberg Y. Neural word embedding as implicit matrix factorization//Proceedings of the Advances in Neural Information Processing Systems 27. Montreal, Canada, 2014: 2177-2185

[78] Church K W, Hanks P. Word association norms, mutual information, and lexicography//Proceedings of the 27th Annual Meeting on Association for Computational Linguistics. Stroudsburg, USA, 1989: 76-83

[79] Li Y, Xu L, Tian F, et al. Word embedding revisited: A new representation learning and explicit matrix factorization perspective//Proceedings of the 24th International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 3650-3656

[80] Shi T, Liu Z. Linking GloVe with Word2vec. arXiv preprint arXiv: 1411.5595, 2014

[81] Suzuki J, Nagata M. A unified learning framework of skip-grams and global vectors//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. Beijing, China, 2015: 186-191

[82] Turian J, Ratinov L, Bengio Y. Word representations: A simple and general method for semi-supervised learning//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA, 2010: 384-394

[83] Baroni M, Dinu G, Kruszewski G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, 2014: 238-247

[84] Milajevs D, Kartsaklis D, Sadrzadeh M, et al. Evaluating neural word representations in tensor-based compositional settings//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 708-719

[85] Finkelstein L, Gabrilovich E, Matias Y, et al. Placing search in context: The concept revisited. ACM Transactions on Information Systems, 2002, 20(1): 116-131

[86] Hill F, Reichart R, Korhonen A. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. Computational Linguistics, 2015, 41(4): 665-695

[87] Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics. Atlanta, USA, 2013: 746-751

[88] Arora S, Li Y, Liang Y, et al. Random Walks on Context Spaces: Towards an Explanation of the Mysteries of Semantic Word Embeddings. arXiv preprint arXiv: 1502.03520, 2015

[89] Tjong Kim Sang E F, Buchholz S. Introduction to the CoNLL-2000 shared task: Chunking//Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th Conference on Computational Natural Language Learning. Stroudsburg, USA, 2000: 127-132

[90] Fan R-E, Chang K-W, Hsieh C-J, et al. LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research, 2008, 9: 1871-1874

[91] Schnabel T, Labutov I, Mimno D, et al. Evaluation methods for unsupervised word embeddings//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 298-307

[92] Landauer T K. On the computational basis of learning and cognition: Arguments from LSA. Psychology of Learning and Motivation, 2002, 41: 43-84

[93] Ling W, Dyer C, Black A W, et al. Two/Too simple adaptations of Word2Vec for syntax problems//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado, 2015: 1299-1304

[94] Guo J, Che W, Wang H, et al. Revisiting embedding features for simple semi-supervised learning//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 110-120

[95] Tian F, Dai H, Bian J, et al. A probabilistic model for learning multi-prototype word embeddings//Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland, 2014: 151-160

[96] Neelakantan A, Shankar J, Passos A, et al. Efficient non-parametric estimation of multiple embeddings per word in vector space//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1059-1069

[97] Chen X, Liu Z, Sun M. A unified model for word sense representation and disambiguation//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 1025-1035

[98] Miller G A. WordNet: A lexical database for English. Communications of the ACM, 1995, 38(11): 39-41

[99] Qiu L, Cao Y, Nie Z, et al. Learning word representation considering proximity and ambiguity//Proceedings of the 28th AAAI Conference on Artificial Intelligence. Québec, Canada, 2014: 1572-1578

[100] Liu Y, Liu Z, Chua T-S, et al. Topical word embeddings//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin Texas, USA, 2015: 2418-2424

[101] Liu P, Qiu X, Huang X. Learning context-sensitive word embeddings with neural tensor skip-gram model//Proceedings of the 24th International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 1284-1290

[102] Li J, Jurafsky D. Do multi-sense embeddings improve natural language understanding?//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 1722-1732

[103] Botha J A, Blunsom P. Compositional morphology for word representations and language modelling//Proceedings of the 31st International Conference on Machine Learning. 2014: 1899-1907

[104] Qiu S, Cui Q, Bian J, et al. Co-learning of word representations and morpheme representations//Proceedings of the 25th International Conference on Computational Linguistics. Dublin, Ireland, 2014: 141-150

[105] Sun F, Guo J, Lan Y, et al. Inside out: Two jointly predictive models for word representations and phrase representations// Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016: 2821-2827

[106] Ling W, Dyer C, Black A W, et al. Finding function in form: Compositional character models for open vocabulary word representation//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 1520-1530

[107] Rubinstein D, Levi E, Schwartz R, et al. How well do distributional models capture different types of semantic knowledge?//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. Beijing, China, 2015: 726-730

[108] Baker C F, Fillmore C J, Lowe J B. The Berkeley Framenet project//Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics. Montreal, Canada, 1998: 86-90

[109] Yu M, Dredze M. Improving lexical embeddings with semantic knowledge//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, 2014: 545-550

[110] Bian J, Gao B, Liu T-Y. Knowledge-Powered Deep Learning for Word Embedding. Calders T, Esposito F, Hüllermeier E, et al. Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2014, 8724: 132-148

[111] Xu C, Bai Y, Bian J, et al. RC-NET: A general framework for Incorporating knowledge into word representations// Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. Shanghai, China, 2014: 1219-1228

[112] Liu Q, Jiang H, Wei S, et al. Learning semantic word embeddings based on ordinal knowledge constraints// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. Beijing, China, 2015: 1501-1511

[113] Faruqui M, Dodge J, Jauhar S K, et al. Retrofitting word vectors to semantic lexicons//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado, 2015: 1606-1615

[114] Griffiths T L, Steyvers M, Tenenbaum J B. Topics in semantic representation. Psychological Review, 2007, 114(2): 211-244

[115] Schunn C D. The presence and absence of category knowledge in LSA//Proceedings of the 21st Annual Conference of the Cognitive Science Society. Vancouver, Canada, 1999: 643-648

[116] Attwell D, Laughlin S B. An energy budget for signaling in the grey matter of the brain. Journal of Cerebral Blood Flow & Metabolism, 2001, 21(10): 1133-1145

[117] Olshausen B A, Field D J. Sparse coding with an overcomplete basis set: A strategy employed by V1? Vision Research, 1997, 37(23): 3311-3325

[118] Vinson D P, Vigliocco G. Semantic feature production norms for a large set of objects and events. Behavior Research Methods, 2008, 40(1): 183-190

[119] Murphy B, Talukdar P, Mitchell T. Learning effective and interpretable semantic models using non-negative sparse embedding//Proceedings of the 24th International Conference on Computational Linguistics. Mumbai, India, 2012: 1933-1950

[120] Faruqui M, Tsvetkov Y, Yogatama D, et al. Sparse overcomplete word vector representations//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. Beijing, China, 2015: 1491-1500

[121] Sun F, Guo J, Lan Y, et al. Sparse word embeddings using ℓ_1 regularized online learning//Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, USA, 2016: 2915-2921

[122] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization. Nature, 1999, 401(6755): 788-791

[123] Turney P D. Similarity of semantic relations. Computational Linguistics, 2006, 32(3): 379-416

[124] Tsvetkov Y, Faruqui M, Ling W, et al. Evaluation of word vector representations by subspace alignment//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 2049-2054

[125] Chen Y, Perozzi B, Al-Rfou R, et al. The expressive power of word embeddings//Proceedings of the ICML 2013 Workshop on Deep Learning for Audio, Speech, and Language Processing. Atlanta, USA, 2013: 1-11

[126] Faruqui M, Dyer C. Non-distributional word vector representations//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. Beijing, China, 2015: 464-469

[127] Tang D, Wei F, Qin B, et al. Building large-scale twitter-specific sentiment lexicon: A representation learning approach// Proceedings of the 25th International Conference on Computational Linguistics. Dublin, Ireland, 2014: 172-182

[128] Chen X, Xu L, Liu Z, et al. Joint learning of character and word embeddings//Proceedings of the 242th International Joint Conference on Artificial Intelligence. Buenos Aires, Argentina, 2015: 1236-1242

[129] Li Y, Li W, Sun F, et al. Component-enhanced Chinese character embeddings//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 829-834

[130] Wang H, Li C, Zhang Z, et al. Topic modeling for short texts with auxiliary word embeddings//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. Pisa, Italy, 2016: 165-174

[131] Ganguly D, Roy D, Mitra M, et al. Word embedding based generalized language model for information retrieval// Proceedings of the 38th International ACM SIGIR Conference

on Research and Development in Information Retrieval. Santiago, Chile, 2015: 795-798

[132] Vulić I, Moens M-F. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings //Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. Santiago, Chile, 2015: 363-372

[133] Mitra B. Exploring session context using distributed representations of queries and reformulations//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. Santiago, Chile, 2015: 3-12

[134] Grbovic M, Djuric N, Radosavljevic V, et al. Context- and content-aware embeddings for query rewriting in sponsored search//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. Santiago, Chile, 2015: 383-392

[135] Fu R, Guo J, Qin B, et al. Learning semantic hierarchies via word embeddings//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, 2014: 1199-1209

[136] Soricut R, Och F. Unsupervised morphology induction using word embeddings//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado, 2015: 1627-1637

[137] Fyshe A, Talukdar P P, Murphy B, et al. Interpretable semantic vectors from a joint model of brain- and text- based meaning//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, Maryland, 2014: 489-499

[138] Faruqui M, Dyer C. Improving vector space word representations using multilingual correlation//Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden, 2014: 462-471

[139] Wilson B J, Schakel A M J. Controlled Experiments for Word Embeddings. arXiv preprint arXiv: 1510.02675, 2015

[140] Schakel A M J, Wilson B J. Measuring Word Significance using Distributed Representations of Words. arXiv preprint arXiv: 1508.02297, 2015

[141] Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 1631-1642

[142] Socher R, Huang E H, Pennin J, et al. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection// Proceedings of the Advances in Neural Information Processing Systems 24. Granada Spain, 2011: 801-809

[143] Blacoe W, Lapata M. A comparison of vector-based representations for semantic composition//Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing. Jeju Island, Korea, 2012: 546-556

[144] Mitchell J, Lapata M. Composition in distributional models of semantics. Cognitive Science, 2010, 34(8): 1388-1429

[145] Gershman S J, Tenenbaum J B. Phrase similarity in humans and machines//Proceedings of the 37th Annual Conference of the Cognitive Science Society. Pasadena, USA, 2015: 776-781

[146] Chandar A P S, Lauly S, Larochelle H, et al. An autoencoder approach to learning bilingual word representations. Advances in Neural Information Processing Systems 27. Curran Associates, Inc., 2014: 1853-1861

[147] Erk K. Representing words as regions in vector space// Proceedings of the 13th Conference on Computational Natural Language Learning. Stroudsburg, USA, 2009: 57-65

[148] Vilnis L, McCallum A. Word representations via Gaussian embedding//Proceedings of the International Conference on Learning Representations. San Diego, USA, 2015: 1-12

[149] Koo T, Carreras X, Collins M. Simple semi-supervised dependency parsing//Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics. Columbus, Ohio, 2008: 595-603

[150] Liang P. Semi-Supervised Learning for Natural Language [M. S. disseratation]. Massachusetts Institute of Technology, Cambridge, USA, 2005

SUN Fei, Ph. D. candidate. His research interests mainly focus on representation learning and text mining.

GUO Jia-Feng, Ph. D., associate professor. His research focuses on information retrieval and data mining.

LAN Yan-Yan, Ph. D., associate professor. Her research interests include machine learning, learning to rank and information retrieval.

XU Jun, Ph. D., professor. His research interests include information retrieval and data mining.

CHENG Xue-Qi, Ph. D., professor. His research interests include network science, network and information security, Web search and data mining.

Background

Word representation is a fundamental problem in natural language processing. Traditional one-hot representations suffer from the data sparsity in practice due to missing semantic relation between words. Different from the one-hot representations, distributed word representations encode the semantic meaning of words as dense, real-valued vectors in a low-dimensional space. As a result, the distributed word representations can alleviate the data sparsity issues. Since the success in neural language model, distributed word representations have been widely used in natural language processing, like language model, machine translation, POS tagging, and sentiment analysis. At the same time, various models are proposed to leaning distributed word representations including matrix factorization and neturon network approaches. In this paper, we combed the development of models for learning distributed word representations. Furthermore, we found that all these models are built on the distributional hypothesis but with different contexts. From this perspective, we can group these models into two classes, syntagmatic and paradigmatic. Then, we summarize the key challenges in learning word representation including polysemous word representations, rare word representations, fine-grained semantic modeling, the Interpretability of distributed word representations, and the evaluation of word representation. In addition, we also discuss the latest and potential solutions for these challenges. At last, we give a future outlook on the research and application directions.

This work was supported by the 973 Program of China under Grant Nos. 2014CB340401 and 2013CB329606, the National Natural Science Foundation of China under Grant Nos. 61232010, 61472401, 61425016, 61203298, and the Youth Innovation Promotion Association CAS under Grant Nos. 20144310 and 2016102.