Word2vec 的核心架构及其应用

熊富林[1],邓怡豪[1],唐晓晟[2]

(1.北京邮电大学信息与通信工程学院,北京 100876)
(2.北京邮电大学 WTI 实验室,北京 100876)

[摘要]　神经网络概率语言模型是一种新兴的自然语言处理算法,该模型通过学习训练语料获得词向量和概率密度函数,词向量是多维实数向量,向量中包含了自然语言中的语义和语法关系,词向量之间余弦距离的大小代表了词语之间关系的远近,词向量的加减代数运算则是计算机在“遣词造句”. 近年来,神经网络概率语言模型发展迅速,Word2vec 是最新技术理论的合集. 首先,重点介绍 Word2vec 的核心架构 CBOW 及 Skip-gram; 接着,使用英文语料训练 Word2vec 模型,对比两种架构的异同; 最后,探讨了 Word2vec 模型在中文语料处理中的应用.

[关键词]　自然语言处理,Word2vec CBOW Skip-gram,中文语言处理
[中图分类号] TP391.1　[文献标志码] A　[文章编号] 1672-1292(2015)01-0043-06

The Architecture of Word2vec and Its Applications

Xiong Fulin[1],Deng Yihao[1],Tang Xiaosheng[2]

(1.School of Information and Communication Engineering,Beijing University of Posts and Telecommunications,Beijing 100876,China)
(2.Wireless Technology Innovation,Beijing University of Posts and Telecommunications,Beijing 100876,China)

Abstract: Word2vec is a combination of neural probabilistic language model,which includes CBOW model and Skip-gram model in terms of architecture. This paper will introduce the technology of Word2vec. Firstly,the paper will elaborate the theory of Word2vec architecture; secondly,an English corpus which is extracted from Wikipedia will be used to train the model,and a set of results will be shown; lastly,the application of Word2vec in the language of Chinese will be explored,a result will also be presented precisely.

Key words: NPL,Word2vec CBOW Skip-gram,Chinese-language-processing

统计语言模型的一般形式是给定已知的一组词,求解下一个词的条件概率,具体如下式所示:

$$\hat{p}(w_1^T)=\prod_{t=1}^{T}\hat{p}(w_t|w_1^{t-1}),\tag{1}$$

其中,w_t 是第 t 个词,w_i^j 为上文词序列:

$$w_i^j=(w_i,w_{i+1},w_{i+2}\cdots,w_{j-1},w_j).$$

统计语言模型的一般形式直观、准确,n 元模型就是其中的一种. 在构建 n 元模型时,有如下假设: 在不改变词语在上下文中的顺序的前提下,距离相近的词关系越近,或者说关联度越大; 而距离较远的词关系越小,当距离足够远时,我们认为词语之间没有关联度(或者说关联度很小,在构建模型时可以忽略). 基于以上认识,我们的模型可以简化为如下模型:

$$\hat{p}(w_t|w_1^{t-1})\approx\hat{p}(w_t|w_{t-n+1}^{t-1}).\tag{2}$$

问题看似解决了,但是,上述模型设定了前提条件: 只有出现在训练语料中的句子才能运用此模型,并且句子中词语的顺序还不能改变. 换句话说,如果一个句子并没有出现在训练语料中,那么按照上述模型,其概率就应该为 0.

有一种解决方案是“拼凑”,即我们从训练语料中取出片段来拼凑出未出现在语料中的句子,这些片段可长可短,1 个词、2 个词等等(如果是 n 个词那么就等同于句子在训练语料中出现了),具体该方法如

收稿日期: 2014-08-16.
通讯联系人: 熊富林,硕士,研究方向: 数据挖掘. E-mail: fulinxiong@gmail.com

何获取条件概率请参考 back-off trigrammodels(Katz,1987),Smoothed trigram models(Jelinek and Mercer,1980). 但是使用这种方法,花费巨大代价也很难取得足够性能的提升(Goodman,2001).

上述模型没有完全利用语料的信息:(1)没有考虑距离更远的词语与当前词的关系,即超出范围 n 的词被忽略了,而这两者很可能是有关系的. 比如,"华盛顿是美国的首都"是当前语句,隔了大于 n 个词的地方又出现了"北京是中国的首都",在 n 元模型中,"华盛顿"和"北京"是没有关系的,然而,两个句子却隐含了语法以及语义关系,即"华盛顿"和"北京"都是名词,并且分别是美国和中国的首都;(2)忽略了词语之间的相似性,即上述模型无法考虑词语的语法关系. 例如,语料中的"鱼在水中游"应该能够帮助我们产生"马在草原上跑"这样的句子,因为在这两句中,"鱼"与"马","水"与"草原","游"与"跑","中"和"上"具有相同的语法特性. 而在神经网络概率语言模型当中,这两种信息将被充分利用到.

1 神经网络概率语言模型的发展

神经网络概率语言模型经历了很长的发展阶段[1-10]. 由 Bengio 等人提出的模型 NNLM(Neural network language model)[1] 最为系统知名,以后的发展工作都参照此模型进行.

历经十余年的研究,神经网络概率语言模型有了很大的发展. 在架构方面,有了比 NNLM 更简单的 CBOW 模型[12,13]、Skip-gram 模型[12,13];其次,在训练算法方面,出现了 Hierarchical Softmax 算法[11]、负采样算法(Negative Sampling)[2,6,10],以及为了减小频繁词对结果准确性和训练速度的影响而引入的欠采样(Subsumpling)技术. 本文将要讨论的是 Word2vec 的架构,即 CBOW 和 Skip-gram.

2 Word2vec 模型的核心架构

在开始之前,我们引入模型复杂度[3],模型复杂度的定义如下:

$$O = E * T * Q. \tag{3}$$

其中,E 表示训练的次数,T 表示训练语料中词的个数,Q 因模型而异. E 值不是我们关心的内容,T 与训练语料有关,其值越大模型就越准确. Q 在下面讲述具体模型时再讨论.

如上节所言,NNLM 模型是神经网络概率语言模型的基础模型. 在 NNLM 模型中,从隐含层到输出层的计算量是主要影响训练效率的地方[1],CBOW 和 Skip-gram 模型考虑去掉隐含层. 尽管神经网络的隐含层很有吸引力,但是实践证明了新架构的可行性. 新架构训练的词向量的精确度可能不如 NNLM 模型,但是这一点可以依靠增加训练语料的方法来完善. 下面,我们介绍 CBOW 和 Skip-gram 模型.

2.1 CBOW 模型

CBOW 模型简单理解就是上下文决定当前词出现的概率. 在 CBOW 模型中,上下文所有的词对当前词出现概率的影响的权重是一样的,因此叫做 CBOW(continuous bag-of-words model)模型,正如在袋子中取词,取出数量足够的词就可以了,至于取出的先后顺序是无关紧要的. CBOW 的模型图如图 1 所示.

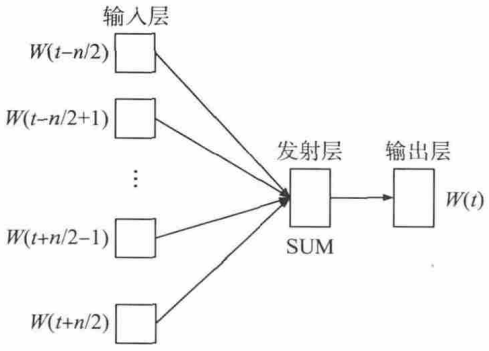

图 1 CBOW 模型示意

Fig. 1 CBOW model

CBOW 模型的训练复杂度为

$$Q = N * D + D * |V|. \tag{4}$$

其中,N 为输入层窗口长度,D 为发射层维度,$|V|$ 为训练语料的词典大小(不同词语的个数). 复杂度就是简单矩阵计算. 主要的计算量为 $D * |V|$.

2.2 Skip-gram 模型

Skip-gram 模型是一个简单但却非常实用的模型. 在自然语言处理中,语料的选取是一个相当重要的问题:第一,语料必须充分. 一方面词典的词量要足够大,另一方面要尽可能多地包含反映词语之间关系的句子,例如,只有"鱼在水中游"这种句式在语料中尽可能地多,模型才能够学习到该句中的语义和语法关系,这和人类学习自然语言一个道理,重复的次数多了,也就会模仿了;第二,语料必须准确. 也就是说所选取的语料能够正确反映该语言的语义和语法关系,这一点似乎不难做到,例如中文里,《人民日报》的

语料比较准确. 但是,更多的时候,并不是语料的选取引发了对准确性问题的担忧,而是处理的方法. 如前文所述,n 元模型中,因为窗口大小的限制,导致超出窗口范围的词语与当前词之间的关系不能被正确地反映到模型之中,如果单纯扩大窗口大小又会增加训练的复杂度. Skip-gram 模型的提出很好地解决了这些问题.

顾名思义,Skip-gram 就是"跳过某些符号",例如,句子"中国足球踢得真是太烂了"有 4 个 3 元词组,分别是"中国足球踢得"、"足球踢得真是"、"踢得真是太烂"、"真是太烂了",可是我们发现,这个句子的本意就是"中国足球太烂",可是上述 4 个 3 元词组并不能反映出这个信息. Skip-gram 模型却允许某些词被跳过,因此可以组成"中国足球太烂"这个 3 元词组. 如果允许跳过 2 个词,即 2-Skip-gram,那么上句话组成的 3 元词组为:

表 1 Skip-gram 示例

Table 1 The example of Skip-gram

Tri-grams	2-Skip-gram-tri-grams
"中国足球踢得"、"足球踢得真是"、"踢得真是太烂"、"真是太烂了"	"中国足球踢得"、"中国足球真是"、"中国足球太烂"、"中国踢得真是"、"中国踢得太烂"、"中国踢得了"、"中国真是太烂"、"中国真是了"、"足球踢得真是"、"足球踢得太烂"、"足球踢得了"、"足球真是太烂"、"足球真是了"、"足球太烂了"、"踢得真是太烂"、"踢得真是了"、"踢得太烂了"、"真是太烂了"

由表 1 可以看出: 一方面,Skip-gram 反映了句子的真实意思,在新组成的这 18 个 3 元词组中,有 8 个词组能够正确反映例句的真实意思; 另一方面,扩大了语料,3 元词组由原来的 4 个扩展为 18 个,如前文所述,语料的扩展能够提高训练的准确度,获得的词向量更能反映真实的文本含义.

Skip-gram 模型的示意图如图 2 所示.

Skip-gram 模型的训练目标就是使得下式的值最大:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leqslant j\leqslant c}\log p(w_{t+j}\mid w_t). \tag{5}$$

其中,c 是窗口的大小,在 Skip-gram 模型中就是指 n-Skip-gram 中的 n 的大小,T 是训练文本的大小. 在 Word2vec 中,使用的是 c-Skip-gram-bi-grams. 基本的 Skip-gram 模型计算条件概率如下式:

$$p(w_O\mid w_I)=\frac{\boldsymbol{v}_{w_O}^{T}\boldsymbol{v}_{w_I}}{\sum_{w=1}^{|V|}\exp(\boldsymbol{v}_{w}^{T}\boldsymbol{v}_{w_I})}. \tag{6}$$

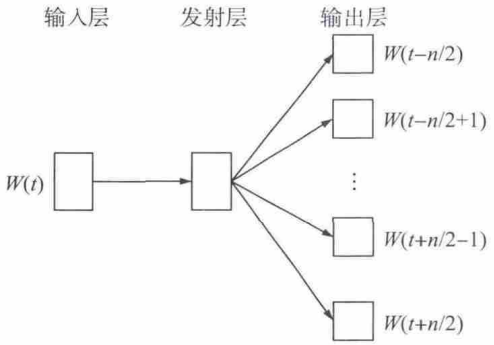

图 2 Skip-gram

Fig. 2 Skip-gram model

其中,$\boldsymbol{v}_w$ 和 $\boldsymbol{v}'_w$ 分别是词 w 的输入和输出向量. 上式是不实用的,因为其计算量与词典大小成正比.

Skip-gram 的计算复杂度为 $Q=c*(D+D*|V|)$,相比 CBOW 模型,其计算复杂度更高.

3 实验结果

本章使用英语语料对模型进行实验验证,第 4 章将探讨 Word2vec 在中文处理中的应用. 在第 3 章中,我们首先比较两种 Word2vec 模型的核心架构.

本文使用 14 组实验对模型进行了检验. 14 组实验包含各种语义和语法实验组合. 实验结果以准确率衡量,包括语义准确率、语法准确率以及平均准确率. 我们使用了两种架构分别进行实验,比较了不同向量维度的实验结果的差别.

14 组不同的实验包括了 5 组语义实验和 9 组语法实验. 14 组实验的示例如表 2.

衡量的方法是使用前 3 个词语预测最后一个词,如果相同,则预测准确; 否则错误,准确预测的数目占实验条目总数的百分比即是其准确率.

在下面的实验中,将采用计算平均准确率的方法,即语义准确率是 5 组语义实验的平均准确率,语法准确率是 9 组语法实验的平均准确率,最后的平均准确率是 14 组实验的平均准确率.

Skip-gram 比 CBOW 的准确率高,代价是耗时很长,CBOW 的训练速度非常快;在一定范围内,维度的增加能大大促进准确率的提升,但是超过范围后准确率不增反降;第 6、7、8 组及 11、12 组实验显示 CBOW 的语义准确率很低;9、10 组显示 Skip-gram 与负采样技术结合的效果很不理想,第 10 组实验耗时最长,其准确率相对于 Skip-gram 最好的结果却下降了;CBOW 与 neg 结合的效果较好,但是对语义准确率的提升没有帮助,语法准确率有明显的提升,比较第 2 组和第 12 组,在耗时大于 Skip-gram+hs 的情况下,其语法准确率却几乎相等,说明 Skip+hs 技术优于 CBOW 架构.

当然,上述结果是在训练数据较小的情况下取得的(如前所述,100 M),当训练数据很大时,使用 Skip-gram 技术就非常耗时了,这个时候考虑采用 CBOW+neg 技术就比较合理.

表 2 实验内容示例

Table 2 The example of test

实验组别	示例
语义 1: capital-common-countries	Beijing China Berlin Germany
语义 2: capital-world	Beijing China Kathmandu Nepal
语义 3: currency	Europe euro Japan yen
语义 4: city-in-state	Chicago Illinois Houston Texas
语义 5: family	brother sister father mother
语法 1: adjective to adverb	calm calmly rare rarely
语法 2: opposite	certain uncertain sure unsure
语法 3: comparative	bad worse big bigger
语法 4: superlative	bad worst big biggest
语法 5: present-participle	code coding dance dancing
语法 6: nationality	China Chinese Japan Japanese
语法 7: past-tense	dancing danced flying fell
语法 8: plural	bird birds cow cows
语法 9: plural-verbs	eat eats walk walks

表 3 14 组实验结果

Table 3 The test results

架构 CBOW/Skip-gram	维度	算法 hs/neg	时间/m	语义准确率/%	语法准确率/%	平均准确率/%	编号
Skip-gram	50	hs	4.7	29.2	27.4	27.7	1
	100		7.75	40.9	36.8	37.5	2
	300		21.75	50.0	41.3	42.9	3
	400		28.4	48.3	41.5	42.7	4
	600		42.0	45.2	41.02	41.8	5
CBOW	100	hs	1.8	22.8	28.0	27.1	6
	300		4.5	25.6	32.8	31.5	7
	400		6.0	24.3	32.8	31.2	8
Skip-gram	300	neg-10	28.7	21.0	32.2	30.1	9
		neg-20	55.2	24.0	34.6	32.7	10
CBOW	300	neg-20	9.6	24.1	35.7	33.5	11
	300	neg-30	13.7	23.3	36.6	34.1	12

注:有关算法不是本文讨论的内容,请参考相关文献.

4 探讨 Word2vec 技术在中文处理中的应用

中文和英文有很大的不同:在语义方面,英语重结构,汉语重语义,例如,英语有严格的过去时、现在时和将来时,而汉语则通过上下文来区分;在语法方面,英语多长句、从句,汉语多短句、分句;在编码方面,英语字母和汉字编码也不同,在 UTF-8 编码方式中,英文字母和汉字的编码长度不同.

在前面的讨论中,我们知道,建立神经网络自然语言处理模型时并不考虑语言的语法结构和语义关系,可以将未经训练的神经网络模型理解为一个“刚出生的婴儿”,能够模仿学习任何一种语言;另外,编码长度不同更不会对模型有任何影响,只是在编写代码实现时需要考虑编码的识别和存储. 基于以上两点认识,我们认为 Word2vec 可以用在中文语言处理当中.

为了验证上述想法,我们使用中文语料训练和检验模型. 实验采用的语料来源于 Sohu 新闻网站保存的大量经过手工整理与分类的新闻语料与对应的分类信息[14],其分类体系包括几十个分类节点,网页规模约为 10 万篇文档. 我们选取了其中主题为 IT 类的语料,总共为 40 万余篇文稿. 对这 40 万余篇文稿,首先去除标点符号、回车符号以及其他各类标记,只保留了由汉字构成的词语,接着使用 jieba 分词器作分词处理;阿拉伯数字转为汉字拼写,拼写统一采用汉语小写,即一、二、三……. 最后的训练语料只有一行,从开头第一个词语开始,词语之间使用单个空格符号隔开. 处理后的语料大小 1.9 GB 大小,示例如下所示:

技术规范 草案 之后 推出 的 产品 然而 和 等 厂商 都 宣布 开始 销售 兼容 标准 前 的 产品
但是 在 这个 标准 最终 获得 批准 之前 哪 厂商 也 不能 销售 兼容 这个 标准 的 产品 赛迪网
讯 外电 消息 英特尔 首席 执行官 保罗 奥 特利尼 星期三 披露 了 移动 个人 计算机 这种 计算机.

4.1 中文训练结果演示

首先将训练结果作一个演示,表 4 演示了近似词特性,表 5 演示了语义和语法关系.

表 4 中文语料训练结果,近似词特性演示

Table 4 The Chinese training result and the demo of the saurus feature

原始词	近似词	余弦距离	原始词	近似词	余弦距离	原始词	近似词	余弦距离	原始词	近似词	余弦距离
新华网	广播网	0.814 443	淘宝网	易趣	0.606 710	百度	搜索引擎	0.697 913	中国移动	中国联通	0.731 616
	中青网	0.811 848		开店	0.586 946		搜索	0.693 204		中国电信	0.676 837
	日讯	0.807 077		易趣网	0.573 955		腾讯	0.600 235		联通	0.609 840
	东方网	0.805 755		淘宝	0.541 533		门户网站	0.587 623		中国网通	0.607 076
	搜狐网	0.763 160		狂欢节	0.539 122		雅虎	0.583 331		号段	0.522 171

表 5 中文语料处理结果,语义语法关系示例

Table 5 The demo of semantic and grammatical relation

词 1	词 2	词 3	预测结果	词 1	词 2	词 3	预测结果
中国移动	王建宙	中国联通	常小兵	四川	成都	辽宁	沈阳
百度	李彦宏	腾讯	马化腾	四川	成都	广东	广州
女生	男朋友	男生	谈恋爱	四川	成都	陕西	武汉
妈妈	女人	爸爸	男人	陕西	西安	四川	成都
手机	通讯	电视	新闻频道	韩国	三星	日本	松下
百度	搜索	淘宝	开店	百度	腾讯	大战	登场

4.2 中文处理应用的进一步探讨

本小节探讨 Word2vec 在中文处理应用中的特性,包括两个方面: (1) 维数对训练效果的影响; (2) CBOW 和 Skip-gram 架构在中文处理应用中的异同.

首先,训练语料. 如前面阐述,这些训练语料选自新闻资料,预处理后总大小为 1.9 GB.

其次,验证实验的设计和实现. 本文采用 Word2vec 研究中惯用的做法: (1) 人工标注了 8 000 多组近义词,示例如下:

武侠片　武侠
平　息　平抚
赛　事　大赛
能　耗　电耗.

(2) 验证的方法是: 利用训练的模型预测词组中第一个词的近义词,如果预测结果中前 3 个词包含词组中第二个词,那么预测准确; 否则,预测错误. 最后,统计 8 000 多组近义词组的预测准确率.

最后,实验实施: (1) 训练维度从 10 到 1 000(Skip-gram 为 670); (2) 对 CBOW 和 Skip-gram 分别进行实验,实验历时 5 d.

实验的结果如下,在表 6 中,展示了实验结果的典型值; 在图 3 中,展示了 CBOW 和 Skip-gram 两种架构在不同维度的表现.

表 6 实验结果典型值

Table 6 The typical results

架构	维度				
	10	100	300	500	1 000
CBOW	0.037 841	0.403 226	0.452 233	0.454 715	0.442 928
Skip-gram	0.082 506	0.554 591	0.622 829	0.605 459	—

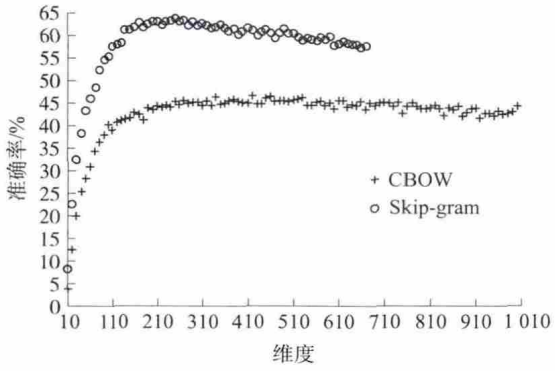

图 3 训练结果

Fig. 3 The training result

由实验结果分析: (1) 使用 CBOW 架构,准确率在维度达到 270 以后趋于平稳; 使用 Skip-gram 架构,准确率在维度达到 170 以后到达峰值水平,峰值出现在维度为 250 的时候; (2) Skip-gram 架构的准确率明显比 CBOW 架构的高,在峰值水平,Skip-gram 架构的准确率比 CBOW 的准确率高 20 个百分点.

综合以上的实验结果,将 Word2vec 应用在中文处理中,维度在 250 左右为宜,维度再大将影响训练的速度; Skip-gram 的准确率明显比 CBOW 高.

5 总结

本文引入Word2vec技术,Word2vec是目前自然语言处理研究的热点技术.

在理论部分,我们介绍了Word2vec技术,并论述了其来龙去脉. Word2vec是神经网络概率语言模型发展到现在的一个技术的集合,包括了模型最新的架构以及最新的算法. 本文重点阐述了模型的核心架构.在模型的验证部分,我们使用英语语料训练Word2vec模型,比较了CBOW和Skip-gram两种核心架构的异同.最后,我们试着探讨了Word2vec技术在中文处理中的应用. 验证了Word2vec技术在中文处理中应用的可行性;探讨了Word2vec技术在中文处理应用中的特性. Word2vec在中文中的应用依然面临难题,其一就是语料的选取,中文,尤其是中文词汇的繁杂给语料的选取带来了麻烦. 这方面,语言学家的帮助将会大有裨益.

Word2vec技术的应用和发展依然有很多工作要做. 希望有更多的人学习、研究及运用Word2vec相关技术.

[参考文献](References)

[1] Bengio Y,Ducharme R,Vincent P. A neural probabilistic language model [J]. Journal of Machine Learning Research,2003,3(7):1 137-1 155.

[2] Michael U G,AapoHyv̈rinen. Noise-contrastive estimation of unnormalized statistical models,with applications to natural image statistics [J]. The Journal of Machine Learning Research,2012,13(2):307-361.

[3] Tomas M,Chen K,Corrado G. Efficient estimation of word representations in vector space [EB/OL].(2013-08-18) [2013-09-07] http://arxiv.org/abs/1301.3781.

[4] Bengio Y,LeCun Y. Scaling Learning Algorithms Towards AI [M]//Large-Scale Kernel Machines. Cambridge:MIT Press,2007.

[5] Mikolov T,Karafi M,Burget L,et al. Recurrent neural network based language model [C]//Proceedings of Interspeech. Chiba,Japan:MIT Press,2010:131-138.

[6] Mikolov T,Ilya S,Kai C,et al. Distributed representations of words and phrases and their compositionality [EB/OL]. [2013-10-16] http://arxiv.org/abs/1310.4546.

[7] Elman J. Finding structure in time [J]. Cognitive Science,1990,14(7):179-211.

[8] Rumelhart D E,Hinton G E,Williams R J. Learning internal representations by back-propagating errors [J]. Nature,1986,323(9):533-536.

[9] 李雷. 基于人工智能机器学习的文字识别方法研究 [D]. 成都:电子科技大学机械电子工程学院,2013.
Li Lei. Character recognition research based on artificial intelligence and machine learning [D]. Chengdu:School of Mechatronics Engineering of University of Electronics Science and Technology of China,2013.(in Chinese)

[10] Andriy M,Yee W T. A fast and simple algorithm for training neural probabilistic language models [EB/OL].(2009-10-12) [2012-06-10] http://arxiv.org/ftp/arxiv/papers/12061.

[11] Frederic M,Yoshua B. Hierarchical probabilistic neural network language model [C]//Proceedings of the International Workshop on Artificial Intelligence and Statistics. Barbados:MIT Press,2005:246-252.

[12] Mikolov T,Kopeck J,Burget L,et al. Neural network based language models for highly inflective languages [C]//Proc. ICASSP. Taipei:ICA,2009:126-129.

[13] Hinton G E,McClelland J L,Rumelhart D E. Distributed Representations [M]//Parallel Dis-Tributed Processing:Explorations in the Microstructure of Cognition. Cambridge:MIT Press,1986.

[14] 许炎,金芝,李戈,等. 基于多Web信息源的主体概念网络获取 [J]. 计算机研究与发展,2013,50(9):1 843-1 854.
Xu Yan,Jin Zhi,Li Ge,et al. Acquiring topical concept network from multiple Web information sources [J]. Journal of Computer Research and Development,2013,50(9):1 843-1 854.(in Chinese)

[责任编辑:黄 敏]