

面向未登录词及多义词的共现性词嵌入改进

李保珍, 顾秀莲

(南京审计大学 信息工程学院, 江苏 南京 211815)

摘要:基于语料库构建词语语义性向量的词嵌入模型,可以定量刻画词语的上下文语义。然而,传统的词嵌入模型在揭示一词多义词汇的语义时,存在着语义空间向量维度不确定或缺乏直观可解释性等局限,此外,对于词汇表外未登录新词语的语义性嵌入识别,尚缺乏有效的途径。针对一词多义问题和未登录词问题,可将词嵌入的优势和词共现的优势相融合,以弥补传统词嵌入模型的语义空间维度不确定、语义维度不可解释及未登录词忽略等方面的不足。主要创新工作包括:基于训练后的词嵌入矩阵与单词归一化的共现矩阵,构建全局性语料词向量;为未登录词创建语料词向量,并与全局性语料词向量进行权重融合,以提高词嵌入的精确率。通过公开数据集的两项实验结果表明,基于词共现的一词多义及未登录词嵌入模型,可有效提升词嵌入的精确度,并可缩短词嵌入的进程时间。

关键词:词嵌入;未登录词;多义词;共现矩阵;词向量

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2022)12-0117-06

doi:10.3969/j.issn.1673-629X.2022.12.018

Co-occurrence Word Embedding Improvement for Unknown and Polysemous Words

LI Bao-zhen, GU Xiu-lian

(School of Information Engineering, Nanjing Audit University, Nanjing 211815, China)

Abstract: The word embedding model of word semantic vector based on corpus can quantitatively describe the context semantics of words. However, the traditional word embedding model has some limitations in revealing the semantics of polysemy words, such as uncertain semantic space vector dimension or lack of intuitive interpretability. In addition, there is still a lack of effective way for the semantic embedding recognition of new words that are not registered outside the vocabulary. Aiming at the problem of polysemy and unlisted words, the advantages of word embedding and word co-occurrence can be combined to make up for the shortcomings of the traditional word embedding model, such as uncertain semantic space dimension, unexplainable semantic dimension and ignoring unlisted words. The main innovative work in this paper includes: constructing global corpus word vector based on the trained word embedding matrix and word normalized co-occurrence matrix; creating a corpus word vector for unregistered words and fusing the weight with the global corpus word vector to improve the accuracy of word embedding. Two experiments on public data sets show that the polysemy and unregistered word embedding model based on word co-occurrence can effectively improve the accuracy of word embedding and shorten the process time of word embedding.

Key words: word embedding; unknown words; polysemous word; co-occurrence matrix; word vector

0 引言

词嵌入是将自然语言语料库中的词语,以向量的形式映射为计算机可处理的、蕴含词语上下文语义的一种数值化处理技术^[1]。word2vec模型是典型的静态词嵌入方法之一,可把自然语言语料库中的词语表示成统一意义维度的结构化的短向量,把词嵌入到一个可计算的向量空间中^[2]。

传统的词嵌入模型是在神经网络的基础上进行简化,移除隐藏非线性激活层,直接将嵌入层和输出层的Softmax layer连接^[3-4]。然而,传统的词嵌入模型只能从固定的词汇量集合中学习已有的词语。随着社会的发展,衍生了许多已有词汇表外的新词,并且许多已有的词汇也会衍生出新的含义。如“桌面”主要是指桌子的表面,但在部分语境中也会表示电脑或手机的打

收稿日期:2021-12-27

修回日期:2022-04-28

基金项目:国家自然科学基金(71673122,72074117);江苏省社科基金项目(20WTB007);江苏省研究生科研创新项目(KYCX21_1948)

作者简介:李保珍(1975-),男,教授,硕士,CCF会员(32974M),研究方向为网络大数据分析、大数据审计等;通讯作者:顾秀莲(1997-),女,硕士,研究方向为计算机审计、自然语言处理。

开界面。如果在一个任务遇到一个新词语,并且这个词语没有出现在用于训练的文本语料库中,即未登录词^[5]。传统的词嵌入模型尚不能为这个新的词语构建嵌入向量。此外,文本语料库中会存在具有多重含义的词语,例如,词语“骄傲”,既有正面含义表示赞赏和高兴,也有负面含义表示自负和自我膨胀,该词语虽然形式相同,但表达的意思却大相径庭,只有考虑上下文语境才能识别出其恰当的含义。传统的词嵌入模型所构建的向量空间维度的确定存在一定的随机性,如何构建合理的向量空间维度,以嵌入形式表示特定词语在上下文语境下的恰当语义,以克服一词多义所带来的歧义问题,尚需进一步改进相关的词嵌入模型。

针对传统的词嵌入模型所存在的未登录词问题及一词多义问题,该文提出基于词共现的词嵌入改进模型。主要创新点为,通过综合考虑词嵌入与词共现两方面因素,将基于传统词嵌入模型所构建的词嵌入向量,与经过词共现计算所构建的词共现向量融合。单独创建未登录词的嵌入向量,并可综合兼顾更多的语义空间维度,以提高词嵌入的语义识别精度,减少一词多义所带来的歧义性干扰。

1 相关工作

国内外对于传统的静态词嵌入模型有大量的改进。针对词的多义性问题,一种方法是词性标注。例如,sense2vec 是 word2vec 的一个扩展,在预处理步骤中,训练语料库中的所有单词都用词性(POS)标记进行注释,然后学习由单词本身及其 POS 标记组成的标记的嵌入^[6]。通过这种方式,产生了不同的表达方式。另一种是对单词出现的上下文进行聚类或使用额外的资源^[7]。例如,Vector-Space Models 模型即通过聚合单词出现的上下文信息来编码单词的多重含义以及 wordnet 来识别单词的多种含义^[8]。

针对未登录词问题,Wang 等人通过生成代表词之间的依赖信息的词向量,提出了 Charter to Word 模型^[9];Bojanowski 等人通过学习文本和文本中所有字符 n -grams 的表示,然后组合单词中出现的 n -grams 的嵌入来计算单词的嵌入^[10]。

然而,上述传统的词嵌入改进途径尚不能同时解决一词多义和未登录词的词嵌入问题。如何同时兼顾一词多义和未登录词问题,以减少一词多义所带来的歧义性干扰,以及未登录词所带来的冷启动问题,需要融合词语之间更多的相关性信息。此外,与传统的词嵌入模型相比,上述改进途径大多需要更多的模型参数及计算资源,面对海量的语料库和实时性计算要求,如何减少模型的参数以提高计算效率也是词嵌入改进的目标。

该文基于负采样训练的连续词袋训练模型,通过权重矩阵和共现矩阵,单独构建基于语料库的词向量模型,该途径可有效解决一词多义的歧义性干扰以及未登录词的冷启动问题,并可有效提升词嵌入模型的计算效率。

2 模型构建

为解决一词多义的歧义性干扰以及未登录词的冷启动问题,以提升词嵌入的效率,需要在词嵌入过程中兼顾词语之间的相似信息。该文基于负采样和连续词袋(CBOW)的词嵌入模型,通过融合共现矩阵所揭示的词语之间相似信息,提出了传统词嵌入模型的改进途径。

2.1 负采样和连续词袋

经典的神经网络词向量语言模型一般有输入层(词向量)、隐藏层和输出层(softmax 层)三层^[11]。传统的词嵌入模型对神经网络做了如下简化:(1)移除隐藏非线性激活层,直接将嵌入层与输出层的 Softmax layer 连接;(2)忽略序列信息,输入的所有词向量汇总到同一个嵌入层;(3)增加上下文窗口得到词嵌入的 CBOW 模型^[12]。

传统的词嵌入模型中引入负采样不仅能够加速模型计算速度,还保证了模型训练的效果^[13],负采样将预料中一个词串的中心词替换成其他词,将不存在的字串作为负样本,只更新隐藏权重矩阵的一部分。基于负采样(Negative Sampling)和 CBOW 的词嵌入模型,输入 CBOW 的训练语料样本、词向量的维度、CBOW 的上下文大小、步长和负采样个数。所有模型参数和字向量都是随机初始化的。之后,对于每个训练样本负采样出 n 个负例中心词并且梯度迭代过程使用随机梯度上升法。最后,输出与词汇表中每个单词对应的模型参数和所有单词向量。

从数学上讲,这是通过首先从 W_0 中选择适当的行来计算上下文词嵌入的总和来实现的。然后将该向量乘以从 W_1 中选择的几行;其中一行对应于目标词,而其他行对应于随机选择 k 个“噪声”词(负采样)。在应用非线性激活函数之后,通过将该输出与标签向量 $t \in R_k + 1$ 进行比较来计算反向传播误差,该标签向量 $t \in R_k + 1$ 在目标字的位置为 1,对于所有 k 个噪声字为 0。在模型训练完成后,目标词的嵌入是 W_0 中的对应行。

用负采样训练的 CBOW 词嵌入模型可以解释为一个神经网络,它预测一个单词与其他单词的相似性。在训练期间文本中每出现一个单词 w ,对应的二进制向量(在上下文单词 w 的位置为 1,在其他位置为 0)被用作网络的输入,并乘以一组权重 W_0 得到嵌入(与

上下文单词对应的 W_0 中行的总和)。然后将该嵌入乘以另一组权重 W_1 (对应于单词嵌入 Y 的完整矩阵), 以产生网络的输出, 即包含单词 w 与所有其他单词的近似相似性的向量 $S_{W_i} \in R_N$ 。然后, 通过将输出的子集与二进制目标向量进行比较来计算训练误差, 该目标向量用作仅考虑少量随机词时真实相似性的近似值。

2.2 共现矩阵

共现矩阵的行或列作为词向量表示后, 它的统计数值代表了词之间的相似程度^[14]。从语料库文本中可构建词语之间的共现矩阵来揭示词与词之间的相似度。根据语料词构成的共现词向量矩阵, 计算词典中每个词之间的成对相似度矩阵, 归一化后可得到 0 到 1 之间的相似度得分。在相似上下文中, 两个词语嵌入向量之间的余弦相似性应接近 1, 当前词语的窗口大小周围共现词语的数量^[15]。由于语料库的词语数量通常较多, 相似矩阵的计算量会随着词语维度增加而呈指数增长。但是一篇文档的词语数量通常有限, 所以基于文档计算词语共现的共现矩阵具有高维和稀疏的特征^[16]。如果选取少量的 k 个随机词, 它们与目标词的相似性有时会接近 0。因此, 一个词语的嵌入和所有嵌入的矩阵的乘积应导致向量接近该词的真正相似性。

改进的词嵌入模型在输入层依然沿用 CBOW 训练模型和负采样, 将语料库放入模型中训练出权重矩阵 $W_{V \times N}$, 并对语料库中每个单词的出现次数进行累加, 得到共现矩阵 $Co_{V \times V}$ 。对共现矩阵 $Co_{V \times V}$ 做归一化处理, 通过将共现矩阵的每一行除以相应词语在语料库文本中出现的次数 M , 即共现矩阵的归一化, 可得到语料库的词语向量 $S_{V \times V}$, 也即语料文本词语之间的相似信息, 通过将其与训练得到的权重矩阵 $W_{V \times N}$ 相乘, 即:

$$W'_{V \times N} = S_{V \times V} * W_{V \times N} \quad (1)$$

可得到一个新的融入了相似信息的权重矩阵 $W'_{V \times N}$ 。

2.3 未登录词向量

在创建整个语料库文本词向量的同时, 为提升模型的精确度和效率, 可为未登录词单独创建语料词向量 X_n 。再针对未登录词对其他语料文本的影响程度, 通过调节权重 m ($0 < m < 1$), 确定整体语料词语和未登录词语之间的影响程度, 更好地解决文档一词多义的问题。未登录词语料词向量是通过未登录词的语料文本词出现频率求平均值得到的。

未登录词的语料文本词向量的计算方式为未登录词语料库中出现的词 x_i 对出现的次数 M_u 求平均值得到的。

$$X_u = \frac{\sum_{i=1}^{M_u} V_{x_i}}{M_u} \quad (2)$$

同样的, 整体语料词向量是通过整体的语料文本词出现频率求平均值得到的。

$$X_g = \frac{\sum_{i=1}^{M_g} V_{x_i}}{M_g} \quad (3)$$

由此, 可以得到新的语料词向量的公式:

$$S_{V \times V} = [m * X_u + (1 - m) * X_g] \quad (4)$$

并将式(4)带入式(1), 可得到最终的词向量权重矩阵 $W'_{V \times N}$ 。

$$W'_{V \times N} = [m * X_u + (1 - m) * X_g] * W_{V \times N} \quad (5)$$

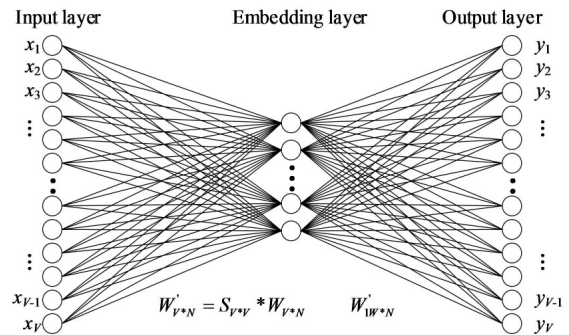


图1 基于负采样和CBOW的共现性词嵌入改进模型

综合上述思路和模型, 可构建面向一词多义及未登录词问题的词嵌入改进模型, 如图1所示, 算法步骤如下:

算法1: 面向一词多义及未登录词的词嵌入改进。

输入: 需要计算词嵌入的文本语料库;

输出: 基于文本语料库的词向量矩阵。

- (1) 整个语料库放入模型中训练出权重矩阵 $W_{V \times N}$;
- (2) 对语料库中每个单词的出现次数进行累加, 得到共现矩阵 $Co_{V \times V}$;
- (3) 对共现矩阵 $Co_{V \times V}$ 做归一化, 得到语料库单词向量 $S_{V \times V}$;
- (4) 将语料词向量 $S_{V \times V}$ 与训练得到的权重矩阵 $W_{V \times N}$ 相乘, 得到一个新的融入了相似信息的权重矩阵 $W'_{V \times N}$;
- (5) 针对未登录词对于其他剩余语料文本的影响程度, 通过调控权重 m ($0 < m < 1$), 决定整体和未登录词之间的影响程度, 以消除词语的歧义。

3 实验

为说明改进的共现性词嵌入模型与传统词嵌入模型相比较的优势, 以传统词嵌入模型作为基线, 通过两个类比任务和命名实体识别任务来评估模型的精确率和 F 值。

3.1 数据准备

类比任务实验采用 text8 数据集及 questions -

words 数据集,数据集来自谷歌官网。其中 text8 数据集是包含空格符和 26 个英文字母的超长句子。questions-words 数据集是包含 14 个类别以及每个类别数据都是 4 列的字符集。

命名实体识别任务实验使用典型数据集 conll2003^[17]。命名实体识别(NER)是信息提取的子任务,conll2003 数据集是由四列数据组成,并用一个空格分隔。数据分为四个类别,LOC(位置)、OGR(组织)、PER(人名)和 MISC。每个单词放在单独的一行,每个句子后面都有一个空行。每行中的第一项是单词,第二项是词性标记,第三项是语法快速标记,第四项是命名实体标记。块标签和命名的实体标签的格式为 I-TYPE,这意味着该单词在 TYPE 类型的短语内。只有当两个相同类型的短语紧跟其后时,第二个短语的第一个单词才会标记为 B-TYPE,以表示它开始了一个新短语。带有标签 O 的单词不是短语的一部分。

数据由三个文件组成:一个训练文件 training 和两个测试文件 Test_a 和 Test_b。第一个测试文件 Test_a 将在开发阶段用于为学习系统找到合适的参数。第二个测试文件 Test_b 用于最终评估。

3.2 实验设置

类比任务用来评测词向量好坏,以此来揭示词嵌入模型在语义关系上的表现^[18]。为了说明使用上下文编码器创建的词嵌入捕获了词语之间有意义的语义和句法关系,在与传统词嵌入模型一起发布的原始分类任务上对它们进行了评估。

首先使用如上所述的负采样训练的 CBOW 词嵌入模型对 text8 语料库进行训练,其中负采样“噪声词”neg 为 13,嵌入维度 d 为 200,上下文窗口大小为 5,随机种子 seed 为 3,从 17 005 207 个单词和 17 006 个句子的语料库中收集 253 854 个独立词,删除频率 c 低于最小计数 min_count 为 5 的单词后,共有 71 290 个唯一单词。71 290 词汇 200 特征训练模型,16 718 844 字训练 2 789.4 s,5 994 字/s。

将传统词嵌入模型在 text8 语料库上进行十次迭代训练以提高精确率,包含约 1 700 万个单词和约 70k 个独特单词的词汇,以及 10 亿个基准数据集的训练部分,其中包含超过 7.68 亿个单词,拥有 486k 个独特单词的词汇量。

命名实体识别任务用作外部评估以说明改进的词嵌入模型相对于传统词嵌入模型的优势。实验将数据集集中的 training 数据通过词嵌入模型训练生成词向量,而其他的如 development 等数据集未登录词表示为零向量。将词向量与逻辑回归分类模型一同训练使用。实体命名识别任务从 218 609 个单词和 946 个句

子的语料库中收集了 20 102 个独特的单词,迭代训练做 20 次。为了验证改进后的模型对于处理未登录词的优势,对权重值 m 从 0 到 1 进行线性增长变化,并对 training 训练集使用传统词嵌入模型进行 100 次迭代作为对比基线词向量。

将 conll2003 数据集集中的 training 用作基于负采样和 CBOW 的词嵌入训练。Test_a 和 Test_b 用作改进词嵌入模型的训练。为了隔离其他因素对性能的影响,实验只使用词语嵌入作为特征信息,不考虑词语大小写或词性标记等其他信息。为了说明将词嵌入与词语的平均上下文词向量相乘可以改进嵌入,改进的共现词嵌入模型仅使用全局上下文词向量进行计算,即 $m = 1$ 。为了说明将词语的平均全局和未登录词语料上下文向量组合作为改进词嵌入的输入,可以有效解决一词多义的问题,对于训练词汇表中出现的词语,设置 $m < 1$ 。

为了评价实验结果,引入精确率和 F 值作为实验评价标准。精确率在文中表示命名实体识别任务中识别正确的样本数与识别总样本的比率。召回率表示识别正确的样本数与实际任务中总样本数的比率。而 F 值表示为精确率和召回率加权调和平均的统计值。

3.3 实验结果及分析

表 1 是类比任务的结果,展示了改进的共现性词嵌入模型和基于负采样和 CBOW 的传统词嵌入模型作为基线,经过迭代 10 次数据集的 14 个类别的精确率对比结果。

整体上来说,10 次迭代结果证明改进的共现性词嵌入模型相比于传统词嵌入模型,精确率高出了约 6.4 个百分点。其中,改进后的 capital-world 和 city-in-state 类别训练的词向量精确性分别是传统词嵌入模型的 1.7 倍和 1.88 倍。改进的共现性词嵌入模型在这些任务类别上表现较好的一个原因可能是,在前四个任务类别中,参与比较的城市和国家名称只有单一的含义。此外,各个类别的精确率尚存在差异,说明词典中的词义多样对精确率有一定的影响,也说明改进模型在词的语义关系获取上有更高的优势。

表 2 展示了 training、Test_a、Test_b 数据训练出的词向量与逻辑回归分类模型应用于 LOC、FER、MISC 及 ORG 4 类命名实体识别任务中的 F 值,其中 m 值设为 0.6。其中,Test_a 训练出的词向量在 LOC、FER、MISC 三项任务中的 F 值最高。Test_a 和 Test_b 训练出的词向量在四项任务的 F 值都高于 training。通过表中实验结果数据的对比,可以看出改进的共现性词嵌入模型训练所得到的词向量,在识别词的多义性性能上优于传统词嵌入模型训练所得到的词向量,抗一词多义带来的歧义干扰能力更强。

表 1 类比任务精确率比对

类别	传统的词嵌入模型(基于负采样和 CBOW 的 word2vec)	改进的共现性词嵌入模型(纳入相似度后)
capital-common-countries	63.8% (323/506)	78.3% (396/506)
capital-world	34.0% (493/1 452)	58.4% (848/1 452)
currency	15.5% (42/268)	19.8% (53/268)
city-in-state	28.6% (449/1 571)	54.0% (849/1 571)
family	79.4% (243/306)	74.5% (228/306)
gram1-adjective-to-adverb	11.0% (83/756)	16.7% (126/756)
gram2-opposite	24.2% (74/306)	24.5% (75/306)
gram3-comparative	64.4% (811/1 260)	64.3% (810/1 260)
gram4-superlative	40.9% (207/506)	38.1% (193/506)
gram5-present-participle	30.5% (303/992)	31.3% (310/992)
gram6-nationality-adjective	70.8% (971/1 371)	67.6% (927/1 371)
gram7-past-tense	30.4% (405/1 332)	31.8% (423/1 332)
gram8-plural	48.9% (485/992)	49.2% (488/992)
gram9-plural-verbs	41.2% (268/650)	32.0% (208/650)
total	42.0% (5 175/12 268)	48.4% (5 934/12 268)

表 2 命名实体识别任务 F 值比对

命名实体识别任务	training	Test _a	Test _b
LOC	61.14	65.75	62.4
MISC	25.62	29.64	28.4
ORG	21.55	24.2	26.95
FER	42.5	45.69	43.93

m 值变化对 F 值的影响如图 2 所示。对 training 数据使用传统词嵌入模型进行 100 次迭代训练, test 数据中的未登录词用零向量表示, 传统词嵌入模型训练出的结果就是 training 数据。实验中对公式中的权重 m 赋值从 0 到 1, 呈线性增长并无限逼近 1。实验结果表明改进的共现性词嵌入模型在命名实体识别任务中性能优于传统的词嵌入模型。当 $m = 0.63$ 时, F 值最

高, 这说明未登录词数据集合与语料库数据集混合得到的词向量优于通过单个数据文本得到的词嵌入向量。改进的共现性词嵌入模型生成的词向量能够更有效地为未登录词创建词向量, 词嵌入向量更符合上下文语境含义, 提高了词嵌入向量的效果。

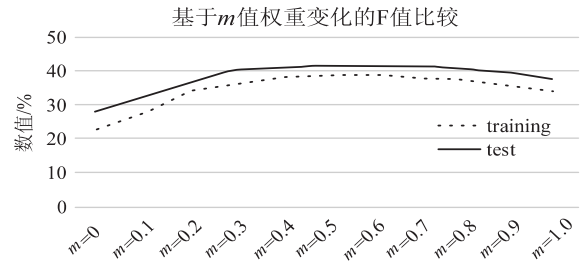
图 2 m 值变化对 F 值的影响

图 3 为基于词嵌入模型三次随机初始化的命名实体识别任务结果。当使用传统词嵌入(虚线)的平均性能被视为基线, 其他嵌入都是词语的全局和局部词向量的各种组合。对比得出, 改进的共现性词嵌入模型在性能上是有所提升的, 全局语料与未登录词语料混合创建出的词向量效果最优。

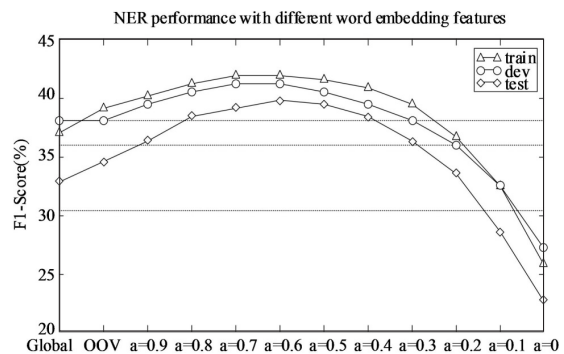


图 3 不同词语嵌入特征下的 NER 性能

4 结束语

综合考虑词嵌入和词共现两方面因素, 提出基于词共现的词嵌入改进模型。通过融合训练后的词嵌入矩阵与词语的共现矩阵, 可在新的词嵌入权重矩阵中蕴含词语间的相似度信息。该途径既可将文本中出现但训练语料库中不存在的词语进行嵌入向量表示, 又可兼顾更多维度的语义空间维度, 进而可更准确识别词语的上下文语境含义, 以减少一词多义所带来的歧义性干扰。因此, 改进的词嵌入模型可以创建未登录词嵌入, 并可根据词语的上下文语境确定其恰当的含义。此外, 改进的词嵌入模型在提升词嵌入处理能力的同时, 也可缩短相关的进程时间, 具有较强的实际意义。未来的研究中将进一步考虑词语间的顺序, 综合考虑词语间的相似性和词语语序, 并将更多维度的信息融入到权重矩阵中, 以进一步提升词嵌入模型的精确率。

参考文献:

- [1] 李 晓,解 辉,李立杰. 基于 Word2vec 的句子语义相似度计算研究[J]. 计算机科学,2017,44(9):256-260.
- [2] LIU Q, LING Z H, JIANG H, et al. Part-of-speech relevance weights for learning word embeddings [J]. arXiv: 1603.07695v1, 2016.
- [3] SANG E, MEULDER F D. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition [C]//Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003. [s. l.]: [s. n.], 2003:142-147.
- [4] 邓晓衡,杨子荣,关培源. 一种基于词义和词频的向量空间模型改进方法[J]. 计算机应用研究,2019,36(5):1390-1395.
- [5] 汪 静,罗 浪,王德强. 基于 Word2Vec 的中文短文本分类问题研究[J]. 计算机系统应用,2018,27(5):209-215.
- [6] 曾 浩,詹恩奇,郑建彬,等. 基于扩展规则与统计特征的未登录词识别[J]. 计算机应用研究,2019,36(9):2704-2707.
- [7] KHOMSAH S. Sentiment analysis on YouTube comments using Word2Vec and random forest[J]. Telematika, 2021, 18(1):61-72.
- [8] XU Z, CHEN B, ZHOU S, et al. A text-driven aircraft fault diagnosis model based on a Word2vec and priori-knowledge convolutional neural network [J]. Aerospace, 2021, 8(4): 112-127.
- [9] JATNIKA D, BIJAKSANA M A, SURYANI A A. Word2-Vec model analysis for semantic similarities in english words [J]. Procedia Computer Science, 2019, 157:160-167.
- [10] 宰新宇,田学东. 基于公式描述结构和词嵌入的科技文档检索方法[J]. 数据分析与知识发现,2020,4(1):131-138.
- [11] 潘 博,于重重,张青川,等. 基于词性与词序的相关因子训练的 word2vec 改进模型[J]. 电子学报,2018,46(8):1976-1982.
- [12] 石隽锋,李济洪,王瑞波. 一种改进的 GloVe 词向量表示学习方法[J]. 中文信息学报,2021,35(4):16-22.
- [13] 巴志超,李 纲,朱世伟. 基于语义网络的研究兴趣相似性度量方法[J]. 现代图书情报技术,2016(4):81-90.
- [14] 李舟军,王昌宝. 基于深度学习的机器阅读理解综述[J]. 计算机科学,2019,46(7):7-12.
- [15] 李一野,邓浩江. 基于改进余弦相似度的协同过滤推荐算法[J]. 计算机与现代化,2020(1):69-74.
- [16] PENG J, CARROLL J, WU Y, et al. Improved word similarity computation for Chinese using sub-word information [C]//Seventh international conference on computational intelligence & security. Sanya:IEEE, 2012:459-462.
- [17] LIAO J, HUANG Y, WANG H, et al. Matching ontologies with Word2Vec model based on cosine similarity [C]//International conference on artificial intelligence and computer vision. [s. l.]:Springer, 2021:367-374.
- [18] 王永贵,郑 泽,李 玥. word2vec-ACV:OOV 语境含义的词向量生成模型[J]. 计算机应用研究,2019,36(6):1623-1628.
- [19] 余 寒,张晋津. 并发加权 μ -演算的一致性内插[J]. 计算机技术与发展,2018,28(11):22-25.
- [20] JIANG J, ZHANG P, MA Z. The μ -calculus model-checking algorithm for generalized possibilistic decision process [J]. Applied Sciences, 2020, 10(7):2594-2608.
- [21] 李前利,江 华. 命题 μ -演算局部模型检测高效算法设计[J]. 计算机工程与应用,2017,53(9):51-56.
- [22] 李 晴. 面向混成系统的形式化建模和性能评价的研究[D]. 南京:南京航空航天大学,2022.
- [23] CLARKE E M, GRUMBERG J O, PELED D A. Model checking[M]. London:The MIT Press, 1999.
- [24] 胡 军,石娇洁,程 桢,等. 一种基于四变量模型的系统安全性建模与分析方法[J]. 计算机科学,2016,43(11):193-199.

(上接第 73 页)

- [5] [M]//Quantitative logic and soft computing 2016. Berlin: Springer, 2017:49-57.
- [6] JING Y, MINER A S. Computation tree measurement language (CTML)[J]. Formal Aspects of Computing, 2018, 30(3):443-462.
- [7] 倪水妹,曹子宁. 面向概率 ZIA 时序及度量性质的检测研究[J]. 小型微型计算机系统, 2015, 36(3):550-555.
- [8] BRADFIELD J, WALUKIEWICZ I. The μ -calculus and model checking[M]//Handbook of model checking. Berlin: Springer, 2018:871-919.
- [9] BRUNI R, MONTANARI U. Temporal logic and the μ -calculus[M]. [s. l.]:Springer International Publishing, 2017.
- [10] 江 华. 命题 μ -演算全局模型检测的高效算法设计[J]. 计算机研究与发展, 2010, 47(8):1424-1433.