文章编号:1000-5862(2021)02-0131-06

一种改进的中文词嵌入模型

杨雨晴,吴水秀*,左家莉
(江西师范大学计算机信息工程学院,江西 南昌 330022)

摘要:针对当前中文词嵌入模型无法较好地建模汉字字形结构的语义信息,提出了一种改进的中文词嵌入模型. 该模型基于词、字和部件(五笔编码)等粒度进行联合学习,通过结合部件、字和词来构造词嵌入,使得该模型可以有效学习汉字字形结构所蕴含的语义信息,在一定程度上提升了中文词嵌入的质量.

关键词:词嵌入;语言模型;自然语言处理
中图分类号:TP 311 文献标志码:A DOI:10.16357/j.cnki.issn1000-5862.2021.02.04

0 引言

在自然语言处理(Natural Language Processing,NLP)领域中,首先要考虑如何将词表示为计算机可理解的形式. 词的表示因而也一直是自然语言处理及相关领域最为关心的问题之一[1].

最为经典词的表示模型是 20 世纪 70 年代由 S. Gerard 等[2]提出的向量空间模型(Vector Space Model,VSM). 向量空间模型的每一维代表 1 个词向量,且仅在该词所表示的那一维上的值非 0,这种 0-1 的高维向量表示方法也被称为独热表示法(One-hot Representation)[2-4]. 独热表示天然地假定词向量之间彼此正交,即词与词相互独立,因而无法表示词与词之间的语义相关性[5]. 此外,独热表示的维度基于字典大小,使得其维数非常高,从而导致“维数灾难”,且由于非 0 值较少,因而会出现数据稀疏现象[6].

潜在语义索引(Latent Semantic Index,LSI)将词-文档矩阵进行奇异值分解,映射到更低维的空间中,得到更为稠密的词的向量表示[1,6]. 通过计算词向量之间的距离(如欧氏距离和向量夹角),可有效表示词之间的相似度,但潜在语义索引在许多自然语言处理任务上的表现仍然不够理想.

自 20 世纪 80 年代起,神经网络技术在机器学习领域中兴起. 文献[7]将神经网络和统计自然语言模型相结合,构造了神经网络语言模型(Neural Network Language Model,NNLM),并提出了词嵌入(Word Embedding)的概念. 词嵌入是词的分布表示[8],以一个稠密的向量表示一个词,其维数远小于词典的大小. R. Collobert 等[9-10]提出 C-W 模型,确定了将词嵌入应用于自然语言处理任务的神经网络框架中,展示了预训练好的词嵌入,可有效提升诸如命名实体识别、机器翻译、文本分类、情感分析等自然语言处理任务的性能[9-11]. 目前,以 Word2vec 为代表的词嵌入已被成功应用于各类自然语言处理任务中[12-14].

近年来,以 ELMO[15]、Bert[16]为代表的预训练模型在多个自然语言任务处理中取得了良好的结果;B. B. Tom 等[17]提出的 GPT-3 在多个自然语言处理任务中也取得了较好结果. 上述这些预训练模型参数数量巨大,通过在大规模语料上的长时间训练,可学习到更好的通用知识,从而有效提升了后续任务的性能[18].

与此同时,如何结合汉字字形结构里蕴含的语义信息,提升中文词嵌入的质量,也引起了研究者的关注. 目前,这方面的工作主要分为 2 类:基于汉字字形组合结构的中文词嵌入研究[19-25]和基于汉字视觉信息的中文词嵌入研究[26-29].

本文从传统的词嵌入模型入手,详细介绍了目前中文词嵌入的研究现状,针对中文词嵌入模型无法较好地建模字形部件的语义信息,提出一种改进的中文词嵌入模型.

收稿日期:2020-09-13
基金项目:国家自然科学基金(60866018)资助项目.
通信作者:吴水秀(1975—),女,江西省南丰人,副教授,主要从事信息检索、中文信息处理和机器学习方面的研究.
E-mail:wushuixiu@jxnu.edu.cn

1 主要的词嵌入模型

传统的独热表示存在“维数灾难”而且难以泛化. 针对上述问题,B. Yoshua 等[7]提出了一种利用神经网络建立统计语言模型的框架,即神经网络语言模型(NNLM).

词嵌入的思想源自 Harris 词的分布假说,即若 2 个词上下文相似,则这 2 个词也是相似的[30-32]. V. Ashish等[33]进一步明确了词的语义是由其上下文决定的.

本质上,词嵌入是词的分布表示[8],以 1 个稠密的向量表示 1 个词,其维数远小于词典大小. 神经网络语言模型还定义了用语言模型训练词嵌入的框架,并被广泛应用于各类预训练模型中[15-18].

1.1 经典的英文词嵌入模型

基于模型推导所采用策略的不同,研究者将上述模型分为 2 类:(i)以 Word2vec[12-13]为代表的模型是基于词局部信息(即词的上下文)的,被称为基于预测的模型;(ii)以 GloVe[14]为代表的模型使用了全局信息,通常是基于全局词的计数信息,因而被称为基于计数的模型[1].

相较于神经网络语言模型,Word2vec 的网络结构更为简单,其网络主体是一个单隐层的前馈神经网络,包含输入层、隐含层和输出层. 因此,Word2vec 的参数更少[5]. 实际上,Word2vec 包含 2 种模型:(i)连续词袋模型(Continuous Bag-of-Words,CBOW),根据给出的上下文预测目标词,输入为目标词的上下文,得到目标词的出现概率;(ii)跳字模型(Skip-Gram,SG),与 CBOW 不同,输入当前词的词向量,输出上下文的词向量.

Word2vec 展示了它无需深的网络也可得到较好的词嵌入[18],而且在词的类比任务上的结果也显示 CBOW 和 SG 所构造的浅层网络结构可有效捕捉潜在的句法和语义相关性[13].

另一个经典的词嵌入模型是全局词向量(GloVe)[14]. 由于 Word2Vec 只包含上下文的局部语义信息,而未考虑全局语义信息,因此会影响词嵌入的表示能力. GloVe 的损失函数要求任意 2 个词向量之间的点积与 2 个词在语料中的共现次数的对数差异最小,从而使得模型可有效结合局部上下文信息和全局词的共现信息.

1.2 主要的中文词嵌入模型

不同于以英语为代表的表音语言系统,汉字自象形文字转变为意音文字,属于表意的文字系统. 因此,汉字的字形结构相应地也蕴含了丰富的语义信息[28-29]. 由于汉字可被分解为更小的部分,如偏旁和部首等字形结构部件,因此这些语义信息也被其字形结构部件所表征.

近年来,研究者试图通过挖掘汉字字形结构里蕴含的语义信息,提升中文词嵌入的质量. 这些方法可以归为 2 类:(i)基于汉字字形组合结构的中文词嵌入模型[19-25],主要有 CWE 模型[19]、MGE 模型[21]、JWE 模型[22]和 SCWE 模型[24]等;(ii)基于汉字视觉信息的中文词嵌入模型,如 GWE 模型[27]、Glyce 模型[28]和 VCWE 模型[29]等.

1.2.1 基于汉字字形组合结构的中文词嵌入模型 Chen Xinxiong 等[19]认为当前的词嵌入模型只考虑了词外部的上下文信息,忽略了词内的语义信息,中文的处理所需的分词使得考虑词内字的语义信息更为必要. 因此,他们提出了 CWE 模型,通过结合字嵌入提升了中文词嵌入的质量. Li Yanran 等[20]认为直接将字嵌入相加得到词嵌入太过简单,他们提出了组合的中文词嵌入模型 CCWE,根据同义词词典得到字之间的语义关系加入词嵌入中.

Yin Rongchao 等[21]在 CWE 模型的基础上提出了一种多粒度的嵌入式 MGE 模型,将上下文分解为词、字和偏旁部首 3 种粒度,引入了偏旁部首所蕴含的语义信息. Yu Jinxing 等[22]提出了联合学习词嵌入式 JWE 模型,通过联合学习词、字和偏旁部首,将上下文的词嵌入、字嵌入和偏旁部首的嵌入求其平均,用于学习目标词的词嵌入.

Shi Xinlei 等[23]利用五笔输入法将汉字分解成若干部分,作为特征来改进中文词的嵌入. Xu Jian 等[24]以跨语言的方式提出了一种基于相似性的字符增强词嵌入式 SCWE 模型,该模型利用翻译工具从其他语言中获得语义知识,挖掘出词与字之间的语义相似性,以获得字的语义信息. Chao Shaosheng 等[25]将词表示为笔画序列,利用笔画 n-grams 信息的学习实现词嵌入.

1.2.2 基于汉字视觉信息的中文词嵌入模型 基于汉字视觉信息的中文词嵌入模型为每个汉字生成给定尺寸的图像,通过卷积神经网络(Convolution Neural Network,CNN)来处理. L. Frederick 等[26]通过生成可视化的字嵌入,提出了一种新的字符级特征,减少了与罕见词相关的数据稀疏性问题. T. R. Su等[27]引入了一种基于像素的 GWE 模型,利用卷积自编码器从图像中学习字符特征.

Sun Chi 等[29]提出了一种基于视觉增强的词嵌

入模型(VCWE)来学习中文词嵌入. 该模型首先使用 CNN 处理汉字的图像获得带有视觉信息的字嵌入,然后再架构双向长短期记忆网络(Bi-directional Long Short Term Memory,Bi-LSTM)[32]和自注意力机制(Self-Attention)[33]将字之间的组合信息加入词嵌入中,最后基于 Skip-Gram 学习词嵌入. 该模型通过 3 层架构综合考虑了字内各部件、字间的组合信息和词的上下文信息.

Meng Yuxian 等[28]提出了一种基于字形向量的词嵌入式 Glyce 模型,它通过挖掘汉字多种格式的历史文本,以获得更为丰富的汉字视觉特征. 考虑到汉字的形成经历了多个阶段,该模型一共选取了金文、隶书、篆书、魏碑、繁体字和简体字等字体,每种字体选用了草书和仿宋 2 种书写风格. 不同于 VCWE,Glyce 采用了多任务学习[10]的方式.

由此可以看出,目前中文词嵌入模型本质上仍是基于 Word2vec 或 GloVe 的框架进行训练,学习汉字字形结构蕴含的语义信息.

然而,在汉字由象形文字发展为音意文字的长期过程中,使得汉字只有部分字形部件用于表意. 如"朝"字是左右结构,它的偏旁部首是"月",在"朝阳"这个词中,"朝"字的表意部件主要是"日"和"月",其含义是"月亮消失后,太阳从草丛里升起",上述的模型若仅考虑偏旁部首"月"的话则语义就会出现偏差. 此外,类似于"巧克力"这样的音译词,单个字的语义对词的语义并无贡献.

因而,目前的方法通过求平均和构建 n-元语言模型等同对待词内的所有字形部件,忽略了字形结构部件对词的语义贡献的有效性问题,可能会将噪声信息带入词嵌入中,使得词的语义出现偏差.

1.3　词嵌入的评价指标

当前评估词嵌入的质量,主要基于词嵌入在一些实际任务中的效果,对词向量的评价方式主要有:(i)基于语言学,以语义相似度任务和类比任务考察词向量的质量;(ii)将词向量应用于各种 NLP 任务中,比如情感分析、词性标注等,考察不同的词向量在这些任务上的效果[13].

2　一种改进的中文词嵌入模型

本文基于 VCWE 模型,提出了一种改进的中文词嵌入模型,基于词、字和部件(五笔编码)等多粒度进行联合学习,结合部件、字和词的语义信息构造词嵌入.

考虑到字内的字形部件的顺序和词内的字的顺序,均会影响词的语义,因此,本文使用了 Bi-LSTM 模型来建模这种顺序结构. 考虑到相同的字形结构部件在不同的词中,其表示的语义信息也不尽相同,因此,本文使用了 Self-Attention 自动学习来解决该问题. 综上所述,本文分别利用 Bi-LSTM 模型和 Self-Attention 构建了字内建模层和词内建模层.

字内建模层学习字内的组合语义信息,将部件作为双向长短期记忆网络的输入,并经过一层自注意力机制层,得到包含部件信息的字嵌入. 词内建模层学习词内字间的组合语义信息,以字嵌入作为双向长短期记忆网络的输入,再经过一层自适应层构建字间的组合语义信息,从而得到含部件信息的词嵌入. 最后,基于 CBOW 框架学习词的上下文的语义信息.

由于本文提出的模型是架构于汉字的字形部件之上,不同于 VCWE 直接对汉字的图像进行处理. 其原因正如 Meng Yuxian 等[28]所言,历史上汉字一直处于演化中,为了更易于书写,现代汉字趋向于更少的笔画,简体汉字可以说是最简化的版本,导致其失去了原有很多重要的象形意义,基于简化汉字的图像,所提取的语义信息不够完整.

2.1　字内建模

给定一个词序列 $D=\{w_1,w_2,\cdots,w_M\}$ 中的词 w_i,其对应的字序列为 $C_i=\{c_1,c_2,\cdots,c_N\}$,其中字 c_j 对应的部件序列为 $R_j=\{r_1,r_2,\cdots,r_K\}$,用 $\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_K$ 分别表示这些部件的嵌入,将部件通过双向长短期记忆网络层得到隐藏状态为 $\boldsymbol{h}_k=(\boldsymbol{h}_k^F;\boldsymbol{h}_k^B)$,$\boldsymbol{h}_k^F=(\boldsymbol{h}_{k-1}^F,\boldsymbol{e}_k)_{\text{LSTM}}$,$\boldsymbol{h}_k^B=(\boldsymbol{h}_{k+1}^B,\boldsymbol{e}_k)_{\text{LSTM}}$,则 $\boldsymbol{H}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n)$,其中 $\boldsymbol{h}_k$ 为第 k 个部件的隐藏状态,LSTM 表示长短期记忆网络. 通过自注意力机制计算注意力向量 $\boldsymbol{\alpha}=\text{softmax}(\boldsymbol{V}\tanh(\boldsymbol{U}\boldsymbol{h}_k^{\mathrm{T}}))$,其中 $\boldsymbol{V}$ 和 $\boldsymbol{U}$ 是可学习的权值参数矩阵,则基于部件信息的字嵌入 $\tilde{\boldsymbol{e}}_j=\sum_{k=1}^{K}\alpha_k\boldsymbol{h}_k$.

2.2　词内建模

得到了基于字的表示后,再将其加入词嵌入中,类似于部件层,也是通过双向长短期记忆网络和自注意力机制获取带有部件信息的词嵌入.

给定词 w_i,其对应的字序列为 $c_1,c_2,\cdots,c_N$,对应的字嵌入为 $\tilde{\boldsymbol{e}}_1,\tilde{\boldsymbol{e}}_2,\cdots,\tilde{\boldsymbol{e}}_N$,将字通过 Bi-LSTM 到达隐藏层 $\tilde{\boldsymbol{h}}_j=(\tilde{\boldsymbol{h}}_j^F;\tilde{\boldsymbol{h}}_j^B)$,$\tilde{\boldsymbol{h}}_j^F=(\tilde{\boldsymbol{h}}_{j-1}^F,\tilde{\boldsymbol{e}}_j)_{\text{LSTM}}$,$\tilde{\boldsymbol{h}}_j^B=(\tilde{\boldsymbol{h}}_{j+1}^B,\tilde{\boldsymbol{e}}_j)_{\text{LSTM}}$.

同样地有 $\tilde{\boldsymbol{H}}=(\tilde{\boldsymbol{h}}_1,\tilde{\boldsymbol{h}}_2,\cdots,\tilde{\boldsymbol{h}}_n)$,其中 $\tilde{\boldsymbol{h}}_j$ 为第 j 个

字的隐藏状态,再用自主力机制得到注意力向量 $\tilde{\boldsymbol{\alpha}} = \mathrm{softmax}(\boldsymbol{V}\tanh(\boldsymbol{U}\tilde{\boldsymbol{h}}_j^{\mathrm{T}}))$,其中 $\boldsymbol{V}$ 和 $\boldsymbol{U}$ 是可学习的权值参数. 则可获得带有部件信息的词嵌入为

$$\boldsymbol{e}_{w_i} = \sum_{j=1}^{N} \tilde{\alpha}_j \tilde{\boldsymbol{h}}_j. \tag{1}$$

再将经过(1)式得到的带有部件信息的词嵌入 $\boldsymbol{e}_{w_i}$ 和使用 CBOW 得到的词向量 $\boldsymbol{e}'_{w_i}$ 进行加权,得到修正后的词向量 $\boldsymbol{m}_{w_i}$:

$$\boldsymbol{m}_{w_i} = \alpha_{i_1}\boldsymbol{e}_{w_i} + \alpha_{i_2}\boldsymbol{e}'_{w_i}. \tag{2}$$

使用式(2)对词嵌入 $\boldsymbol{e}_{w_i}$ 进行加权求和,是因为现代汉语中不可避免会出现大量音译词(如"巧克力"等),则单字的语义对词的语义并无贡献,强行用多粒度的词嵌入模型会带来噪音,导致获得错误的语义信息. 通过加权的方式加大 $\boldsymbol{e}'_{w_i}$ 的权重,可保证其语义更多来自上下文.

2.3 目标函数

本文采用的是 CBOW 框架,该模型的目标函数是最大化

$$L = \sum \log P(w \mid \boldsymbol{c})$$

的对数似然函数,其中 w 是目标词,$\boldsymbol{c}$ 是上下文的词嵌入,代入词向量 $\boldsymbol{m}_{w_i}$,其条件概率为

$$P(w \mid \boldsymbol{c}) = (\exp(\boldsymbol{e}'\boldsymbol{w}^{\mathrm{T}}\boldsymbol{m}_{w_i})) / (\sum \exp(\boldsymbol{e}'\boldsymbol{w}'^{\mathrm{T}}\boldsymbol{m}_{w_i})).$$

2.4 部件建模

五笔编码可有效地表示所有字和词,且相较于笔画或偏旁部首更为简洁,所以,本文选择五笔编码[34]作为字形结构部件.

3 实验

3.1 数据集及预处理

以中文维基百科数据集为实验数据,它包含了 27.8 万篇中文维基百科文章(https://dumps.wikimedia.org/zhwiki),并按如下步骤对该数据集进行了预处理:

(i)使用WikiExtractor工具包(https://github.com/attardi/wikiextractor)将文本数据从".xml"格式转换为".txt"格式;

(ii)使用 opencc 工具包(https://opencc.byvoid.com)将所有字符规范化为简体中文;

(iii)删除非中文字符(如数字和标点符号等);

(iv)使用 THULAC 工具包(https://github.com/thunlp/THULACPython)进行分词;

(v)丢弃出现少于 100 次的单词,最终获得了大小为 66 856 的词汇表;

(vi)使用由 alex_suen 等收集的五笔编码表(https://blog.csdn.net/s_521_h/article/details/42870155)作为字形部件特征;

(vii)选取 CBOW、CWE、JWE 和 VCWE 等 4 个基准模型;为了便于进行比较,选取相同的训练数据集,统一了所有模型的超参数. 其中用于评估的模型维度数设为 100,滑动的上下文窗口大小为 5,负采样的阈值为 10^{-4},每个单词的负样本数为 5,使用 Adam 的小批量异步梯度下降[35],其初始学习率为 0.025.

(viii)测试数据集为基于 Chen Xinxiong 等[19]整理的 Wordsim-240 和 Wordsim-296. Wordsim-240 包含 240 个由人工标注的词对以及各词对相应相似度的打分,Wordsim-296 包含 296 个词对. 在此基础上,还使用了 Xu Jian 等[24]翻译的数据集 CH-MC-30 和 CH-RG-65. CH-MC-30 包含 30 个词对,CH-RG-65 包含 65 个词对. 在训练过程中,在数据集 Wordsim-240 中找到 213 个词对,在数据集 Wordsim-296 中找到 237 个词对,在数据集 CH-MC-30 中找到 26 个词对,在数据集 CH-RG-65 中找到 49 个词对.

3.2 实验结果

表 1 的数据表明 WBWE 模型优于其他基准模型,在 Wordsim-240 数据集中,WBWE 比基准模型得分最高的 VCWE 模型结果更好;在 Wordsim-297 数据集中,WBWE 比其他基准模型的得分稍低;在 CH-MC-30 数据集中,WBWE 得分比最高的 JWE 模型稍高出约 0.23%;在 CH-RG-65 数据集中,比基准模型中得分最高的 CWE 要高出约 5.48%.

表 1 单词相似度任务实验结果对比

模型	Wordsim-240	Wordsim-297	CH-MC-30	CH-RG-65	平均值	正确率提升百分比/%
CBOW	45.10	53.85	43.63	25.88	42.11	-
CWE	45.79	49.76	41.81	41.17	44.63	+2.52
JWE	43.04	51.58	48.09	34.70	44.35	+2.24
VCWE	48.77	54.27	39.16	39.07	45.31	+3.20
WBWE	48.80	50.66	48.32	46.65	48.61	+6.50

注 "+"表示提升了正确率.

为进一步直观理解模型所学到的词嵌入,选取“唐诗”这个词为例,考察各模型所学到的“唐诗”的词嵌入. 通过计算词嵌入之间的余弦相似度度量语义距离,选取不同模型所得到的与“唐诗”语义距离最近的 10 个词,实验结果如表 2 所示.

表 2　以“唐诗”为例各模型所选取的语义距离最近的词

目标词	CBOW	CWE	JWE	VCWE	WBWE
唐诗	宋词	唐末	宋词	古诗	宋词
	词人	唐时	李白	乐府	绝句
	元曲	唐庄	六朝	吟咏	诗风
	欧阳修	唐宋	元曲	李杜	散文
	散曲	唐书	杨慎	王世贞	七言
	陆游	唐朝	楚辞	诗人	乐府
	李白	唐璜	五律	门柱	古诗
	杂剧	唐廷	杜甫	宋诗	诗话
	书法	唐僖宗	黄庭坚	七言	李白
	书画	唐穆宗	离骚	陆游	李商隐

表 2 显示几个模型所选取的大多数词与“唐诗”语义较为相关. 在 CWE 模型的结果中所有词均包含汉字“唐”,其中“唐庄”“唐璜”等词的语义与“唐诗”相去甚远. 主要原因是 CWE 模型字嵌入的比例过大,在计算词嵌入时仅考虑字嵌入的简单组合,忽略了词嵌入的整体语义信息.

在 CBOW、JWE 和 VCWE 模型的结果中包含“杂剧”和“门柱”等,在语义上与“唐诗”并不相关. 此外,还有多个不属于唐代的人物,如“欧阳修”“陆游”“杨慎”“黄庭坚”“王世贞”等,与“唐诗”的语义有一定距离. WBWE 模型所选取的词基本上与“唐诗”的语义较为相关,这表明本文的模型既能充分考虑上下文,也能排除不相关的语义信息.

4　总结与展望

为有效建模汉字字形结构里蕴含的语义信息,目前的大多数中文词嵌入模型是基于汉字的字形组合结构和汉字图形信息的,通过求平均和构建 n-元语言模型等同对待词内的所有字形部件,这些方法会将噪声信息加入词嵌入中.

基于此,本文提出了一种改进的中文词嵌入模型,基于词、字和部件(五笔编码)等多粒度的字符部件,通过构建双层的 Bi-LSTM 和 Self-Attention 以自动学习字形部件所蕴含的语义信息,实验结果显示所提出的模型提升了中文词嵌入的质量.

词嵌入作为自然语言处理任务的底层输入,可直接影响任务的最终结果. 近年来,许多研究发现词嵌入存在明显的性别和种族偏见[36],而这些偏见会在后续自然语言处理任务中被放大,带来严重的后果. 关于英文词嵌入的偏见问题,已有大量的研究展开[37],而中文词嵌入的偏见问题的研究方兴未艾. 将来,希望基于所提出的中文词嵌入模型深入探讨中文词嵌入隐藏的歧视问题.

5　参考文献

[1] Felipe A, Geraldo X. Word embeddings: a survey [EB/OL]. [2019-03-12]. https://arxiv.org/pdf/1901.09069.pdf.

[2] Gerard S, Andrew W, Yang Chungshu. A vector space model for automatic indexing [J]. Communications of the ACM, 1975, 18(11): 613-620.

[3] David D. The most influential paper gerard salton never wrote [EB/OL]. [2019-03-16]. https://www.ideals.illinois.edu/handle/2142/1697.

[4] Peter D T, Patrick P. From frequency to meaning: vector space models of semantics [J]. Journal of Artificial Intelligence Research, 2010, 37(1): 141-188.

[5] Wang Yuxuan, Hou Yutai, Che Wanxiang, et al. From static to dynamic word representations: a survey [EB/OL]. [2019-03-12]. https://link.springer.com/article/10.1007/s13042-020-01069-8.

[6] Scott D, Susan T D, George W F, et al. Indexing by latent semantic analysis [J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407.

[7] Yoshua B, Réjean D, Pascal V, et al. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2003(3): 1137-1155.

[8] Geffrey H, James L M, David E R. Distributed representations [M]. Massachusetts: MIT Press, 1986: 77-109.

[9] Ronan C, Jason W. A unified architecture for natural language processing: deep neural networks with multitask learning [EB/OL]. [2019-03-12]. https://dl.acm.org/doi/10.1145/1390156.1390177.

[10] Ronan C, Jason W, Léon B, et al. Natural language processing (almost) from scratch [J]. Journal of Machine Learning Research, 2011, 12(1): 2493-2537.

[11] Zhang Lei, Wang Shuai, Liu Bing. Deep learning for sentiment analysis: a survey [J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2018, 8(4): e1253.

[12] Tomas M, Chen Kai, Greg C, et al. Efficient estimation of word representations in vector space [EB/OL]. [2019-03-12]. https://arxiv.org/abs/1301.3781v3.

[13] Tomas M, Ilya Sr, Chen Kai, et al. Distributed representations of words and phrases and their compositionality [EB/OL]. [2019-03-12]. https://www.mendeley.com/catalogue/1cc04e87-4750-3f1e-bbd3-7476f9046a47/.

[14] Jeffrey P, Richard S, Christopher D M. Glove: global vectors for word representation [EB/OL]. [2020-02-11]. https://nlp.stanford.edu/pubs/glove.pdf.

[15] Matthew E P, Mark N, Mohit I, et al. Deep contextualized word representations [EB/OL]. [2019-03-12]. https://arxiv.org/pdf/1802.05365.pdf.
[16] Jacob D, Chang Mingwei, Kenton Lee, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2019-03-12]. https://nlp.stanford.edu/seminar/details/jdevlin.pdf.
[17] Tom B B, Benjamin M, Nick R, et al. Language models are Few-Shot Learners [EB/OL]. [2019-03-12]. https://arxiv.org/abs/2005.14165.
[18] Qiu Xipeng, Sun Tianxiang, Xu Yige, et al. Pre-trained models for natural language processing: a survey [EB/OL]. [2019-03-12]. https://arxiv.org/abs/2003.08271v2.
[19] Chen Xinxiong, Xu Lei, Liu Zhiyuan, et al. Joint learning of character and word embeddings [EB/OL]. [2019-03-12]. https://dl.acm.org/doi/10.5555/2832415.2832421.
[20] Li Yanran, Li Wenjie, Sun Fei, et al. Component-enhanced Chinese character embeddings [EB/OL]. [2019-03-12]. https://arxiv.org/abs/1508.06669.
[21] Yin Rongchao, Wang Quan, Li Peng, et al. Multi-granularity Chinese word embedding [EB/OL]. [2019-03-12]. https://www.aclweb.org/anthology/D16-1100.pdf.
[22] Yu Jinxing, Jian Xun, Xin Hao, et al. Joint embeddings of Chinese words, characters, and fine-grained subcharacter components [EB/OL]. [2019-03-12]. http://repository.ust.hk/ir/Record/1783.1-87829.
[23] Shi Xinlei, Zhai Junjie, Yang Xudong, et al. Radical embedding: delving deeper to chinese radicals [EB/OL]. [2019-03-12]. https://www.mendeley.com/catalogue/b7502a9a-cf29-3806-9e84-0120f63fe04b/.
[24] Xu Jian, Liu Jiawei, Zhang Liangang, et al. Improve Chinese word embeddings by exploiting internal structure [EB/OL]. [2019-03-12]. https://www.aclweb.org/anthology/N16-1119.pdf.
[25] Cao Shaosheng, Lu Wei, Li Xiaolong. Cw2vec: learning Chinese word embeddings with stroke *n*-gram information [EB/OL]. [2019-03-12]. http://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/download/14724/14187.
[26] Frederick Liu, Lu Han, Chieh Lo, et al. Learning character-level compositionality with visual features [EB/OL]. [2019-03-12]. https://arxiv.org/abs/1704.04859v1.
[27] Su T R, Lee H Y. Learning Chinese word representations from glyphs of characters [EB/OL]. [2019-03-12]. https://arxiv.org/abs/1704.04859v1.
[28] Meng Yuxian, Wu Wei, Wang Fei, et al. Glyce: Glyph-vectors for Chinese character representations [EB/OL]. [2019-03-12]. https://arxiv.org/pdf/1901.10125.pdf.
[29] Sun Chi, Qiu Xipeng, Huang Xuanjing. VCWE: Visual character-enhanced word embeddings [EB/OL]. [2019-03-12]. https://arxiv.org/pdf/1902.08795.pdf.
[30] Zellig S H. Distributional structure [EB/OL]. [2019-03-12]. https://www.tandfonline.com/doi/pdf/10.1080/00437956.1954.11659520.
[31] Firth J R. A synopsis of linguistic theory, 1930—1955 [EB/OL]. [2019-03-12]. https://www.researchgate.net/publication/238697185_A_synopsis_of_linguistic_theory_1930—1955.
[32] Sepp H, Jürgen S. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.
[33] Ashish V, Noam S, Niki P, et al. Attention is all you need [EB/OL]. [2019-03-12]. https://arxiv.org/abs/1706.03762v5.
[34] **王永民. 数字键汉字编码技术的研究和应用 [J]. 计算机学报, 2008, 31(6): 1046-1055.**
[35] Diederik P K, Jimmy B. Adam: a method for stochastic optimization [EB/OL]. [2019-03-12]. https://arxiv.org/abs/1412.6980v9.
[36] Nikhil G, Londa S, Dan J, et al. Word embeddings quantify 100 years of gender and ethnic stereotypes [EB/OL]. [2019-03-12]. https://arxiv.org/abs/1711.08412.
[37] Wang Tianlu, Xi V L, Nazneen F R, et al. Double-hard debias: tailoring word embeddings for gender bias mitigation [EB/OL]. [2019-03-12]. https://arxiv.org/abs/2005.00965v1.

The Modified Chinese Word Embeddings Model

YANG Yuqing, WU Shuixiu*, ZUO Jiali
(College of Computer and Information Engineering, Jiangxi Normal University, Nanchang Jiangxi 330022, China)

Abstract: Considering that current Chinese word embedding model can not well model the semantic information of Chinese character's glyph structure, an improved Chinese word embedding model is proposed. The model constructs joint learning based on the granularities of words, characters and components (WUBI), can effectively learn the semantic information contained in the Chinese character glyph structure by constructing word embedding with components, characters and words, and improves the quality of Chinese word embeddings.

Key words: word embedding; language model; nature language processing

(责任编辑: 冉小晓)