文章编号：2096-1472(2020)-04-01-06
DOI:10.19644/j.cnki.issn2096-1472.2020.04.001

基于双词语义增强的BTM主题模型研究

王云云，张云华

(浙江理工大学信息学院，浙江 杭州 310018)
✉1528723134@qq.com；605498519@qq.com

摘　要：针对目前短文本在BTM主题模型建模过程中存在的共现双词之间语义联系较弱的问题，提出一种结合cw2vec词向量模型的改进BTM主题模型(cw2vec−BTM)。使用cw2vec模型来训练短文本语料得到词向量，并计算词向量相似度。然后通过设置采样阈值来改进BTM主题模型共现双词的采样方式，增加语义相关词语的被采样概率。实验结果证明，本文提出的改进模型能有效地提高主题模型的主题凝聚度和KL散度。

关键词：短文本；BTM主题模型；词向量；吉布斯采样
中图分类号：TP391.1　**文献标识码**：A

Research on BTM Topic Model Based on Two-Word Meaning Enhancement

WANG Yunyun,ZHANG Yunhua

(*School of Information Science and Technology,Zhejiang Sci-Tech University,Hangzhou* 310018,*China*)
✉1528723134@qq.com;605498519@qq.com

Abstract: Aimingat the problem of weak semantic relationship between co-occurrence words in the short text in the BTM topic model modeling process,an improved BTM topic model (cw2vec-BTM) combined with the cw2vec word vector model was proposed.This research uses the cw2vec model to train short text corpora to obtain word vectors and calculates the word vector similarity.Then by setting the sampling threshold,the sampling method for co-occurrence words in the BTM topic model is improved,while the sampling probability of semantically related words is increased.The experimental results prove that the improved model proposed in this paper can effectively improve the topic cohesion and KL divergence of the topic model.

Keywords:short text;BTM topic model;word vector;gibbs sampling

1 引言(Introduction)

短文本字数少、篇幅小，在分析短文本时，很难准确的挖掘出语义信息。但这些短文本数据反映了人们的日常生活，从中挖掘出有用的信息并应用到实际生活中是非常有意义的。近几年，短文本挖掘的各项研究均取得较好的成果[1−3]。

Blei等人通过LDA模型提取文本的主题信息[4]。但Hong等人指出文档太短不利于训练LDA的情况[5]。针对该问题，Yan等人提出了BTM主题模型来进行短文本建模[6]。BTM通过语料级别词共现来为短文本建模。Zheng等人针对缺少上下文语义信息的问题，提出TF−IWF和BTM融合的短文本分类方法[7]。张芸等人用BTM进行特征扩展，然后用扩展的特征矩阵进行相似度计算[8]。

以上改进的BTM方法都没有考虑到BTM自身存在的双词间缺乏语义联系的问题，因此，本文提出融合cw2vec词向量模型来改变BTM中的共现双词采样方式的改进模型。

2 相关工作(Related work)

2.1 BTM主题模型

BTM是在双词项集层面上对词共现建模，是基于整个语料库的双词来学习文本的主题，即对双词进行建模，构成了双词—主题—词语的三层结构，可以解决短文本稀疏问题。

设有语料库L，语料库L中有一个二元词组集合$|B|$，表示语料中所有的词对，图模型如图1所示。$b=(b_i,b_j)$表示其中的任一词对，b_i、b_j分别表示词对中的词语，z表示词语的主题，K表示主题数目，$z\in[1,K]$，θ表示每篇文档的主题分布，φ表

示不同主题下的词分布，两者皆服从狄利克雷分布。α和β分别是先验参数。

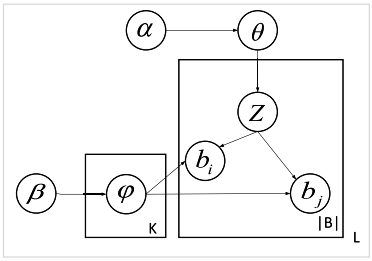

图1 BTM主题模型的图模型结构

Fig.1 Graph model structure of BTM topic model

2.2 cw2vec词向量模型

cw2vec模型是由Cao等人提出的一种基于n元笔画的中文词向量模型[9]，是一种基于skip-gram模型的改进模型，cw2vec模型将笔画信息作为特征，通过使用n-gram笔画来捕捉汉字词语的语义和结构层面的信息。

模型具体介绍如下：

(1)词语分割为字符：为了获取中文字符的笔画信息，将中文词语分割为单个字符。如：大人→大、人。

(2)获取笔画特征：获取中文字符的笔画信息，并将其合并，得到词语的笔画信息。如：大：一丿丶；人：丿丶；大人：一丿丶丿丶。

(3)笔画特征数字化：将中文笔画分为五种不同的类型，用数字代表每一种笔画信息，分别从1到5，如表1所示。

表1 笔画ID对应表

Tab.1 Stroke ID correspondence table

笔画名	横类	竖类	撇类	点类	折类
形，ID	一(㇀)，1	丨(丨)，2	丿(丿)，3	(丶)，4	(乛㇆)，5

所以“大人”这个词语的笔画信息进行特征数字化后可以表示为：

大人：一丿丶 丿丶 → 大人：13434。

(4)动态获取N元笔画特征：提取目标词语笔画信息的n-gram特征。如：

3-gram：134、343、434

4-gram：1343、3434

5-gram：13434

……。

综合以上四个步骤可得n元笔画模型。最后将传统Skip-gram中的词语替换成词语的n-gram笔画特征信息进行训练，以“雾霾治理刻不容缓”为例，cw2vec模型的总体架构如图2所示，当前词语为“雾霾”，上下文词语为“治理”和“刻不容缓”。

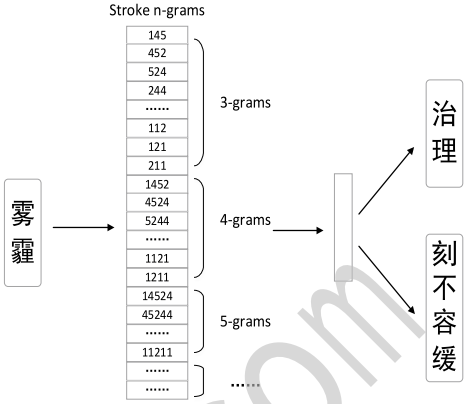

图2 cw2vec模型的总体架构

Fig.2 The overall architecture of the cw2vec model

3 cw2vec-BTM模型(cw2vec-BTM model)

cw2vec-BTM模型的主要改进思想是使用词向量模型训练短文本语料，利用训练出的词向量提取出语料中语义相似的词语，然后结合BTM主题模型参数推理过程，根据词语间的语义相似度阈值，决定出要在词袋中添加的语义相似词语的个数。这样就可以提高相似词语在词袋中被采样到的概率，增强主题的内聚性，以及主题之间的相异性，提高主题模型的主题聚类效果。主要算法模型框架如图3所示。

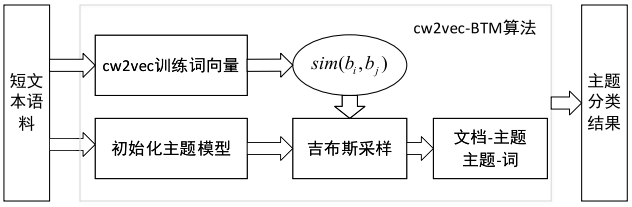

图3 主要算法框架图

Fig.3 Main algorithm framework

该模型主要有以下几个特点：第一，引入cw2vec模型训练词向量，并计算短文本语料中的词向量相似度；第二，改进吉布斯采样方式，添加语义相似度阈值用于模型采样。

3.1 引入cw2vec词向量模型

为了使本文研究的BTM主题模型中的共现双词的语义关联达到最佳，本节对近几年研究的几大词向量模型进行实验对比。

(1)模型介绍及实验设置

实验的训练数据采用中文维基百科训练语料。具体参数设置如表2所示。

表2 模型训练参数设置表

Tab.2 Model training parameter setting table

参数名	参数值
词向量维度	200
窗口大小	5
负采样数目	5
迭代次数	5
最小词频	10
n-gram特征	minn=3,maxn=18
学习率	skip-gram(0.025),CBOW(0.05), CWE(0.05),GWE(0.05), JWE(0.05),cw2vec(0.025)

对比词向量模型有：

2013年，Mikolov等人[10]提出的word2vec实现了两种模型skip-gram和CBOW，应用最为广泛。2015年，Chen等人[11]提出了CWE模型，它是一个基于CBOW模型改进的字符级模型。2017年，Su等人[12]通过图形字符来增强词的表示，提出基于像素的GWE模型。2017年，Yu等人[13]提出了一个联合学习词、字符，以及更加细粒度的部首的方法来学习词向量的模型，称之为JWE模型。2018年，Cao等人[9]提出了基于汉字笔画进行词语信息捕获的cw2vec模型。

(2)实验结果

各模型在中文词语相似度任务上进行了测评，测评的数据集是wordsim-240和wordsim-296，具体实验结果如表3所示。

表3 中文词语相似度结果

Tab.3 Chinese word similarity results

模型	wordsim-240	wordsim-296
skip-gram	50.51	51.95
CBOW	51.04	54.33
CWE	53.19	54.47
GWE	53.23	52.96
JWE	51.85	56.12
cw2vec	55.57	57.07

分析表3，可以得出以下几点结论：

①CWE模型的实验结果总体上要优于skip-gram和CBOW，这是因为word2vec的两种模型是以词为单位进行训练的，相较于融入了字符级信息进行训练的CWE模型，语义表示能力较弱。

②GWE和JWE这两种模型在相似度任务上的表现较不稳定，这可能是由于模型中的像素信息或部首信息是不完整和嘈杂的，影响了模型训练的稳定性。

③cw2vec模型在整体效果上要优于其他几种模型，且是通过构造“n元笔画”和上下文词语之间的相似性函数，直接为每个词语学习单个嵌入。在词语相似度任务上表现最佳。

综上，本文选取cw2vec模型来对语料中的双词进行词向量训练，利用余弦距离来度量词向量的相似度，即b_i和b_j之间的语义距离。如公式(1)所示：

$$\gamma = sim(b_i, b_j) = \frac{b_i \cdot b_j}{|b_i||b_j|} \tag{1}$$

3.2 吉布斯采样方法改进

BTM主题模型是直接对语料库中所有的共现词对进行建模。共现词对就是对文本集语料预处理之后，同一个文档中的任意两个而且无序的词的统称。吉布斯方法直接采样共现双词不利于短文本的主题聚类。

结合上节的词向量相似度结果，本节将利用词向量对主题模型的吉布斯采样方法进行改进。主要是在每次的吉布斯采样中，将采样双词的语义距离γ与语义阈值C进行关系判断，确定是否对双词数量进行扩展，即是否对N_z进行扩展，N_z表示词对b在主题z下被采样到的次数。具体如下：

如果$\gamma > C$，即共现双词的相似度符合要求，则

$$N_z = N_z + \gamma * A \tag{2}$$

其中，A是常量，默认为10；如果不满足要求，则

$$N_z = N_z \tag{3}$$

初始化完成后进行吉布斯采样，如果符合要求，采用公式(4)计算：

$$P(z \mid z_{X_{-b}}, B, \alpha, \beta, \gamma) \propto [n_z + \sum_{K=1}^{K}(\gamma * A) + \alpha] \cdot \frac{[n_{b_1|z} + \sum_{K=1}^{K}(\gamma * A) + \beta][n_{b_2|z} + \sum_{K=1}^{K}(\gamma * A) + \beta]}{[\sum_{b} n_{b|z} + 2\sum_{K=1}^{K}(\gamma * A) + M\beta]^2} \tag{4}$$

不符合要求，则采用公式(5)计算：

$$P(z \mid z_{X_{-b}}, B, \alpha, \beta) \propto [n_z + \alpha] \cdot \frac{[n_{b_1|z} + \beta][n_{b_2|z} + \beta]}{[\sum_{b} n_{b|z} + M\beta]^2} \tag{5}$$

其中，X_{-b}表示语料中不包含词对b的词对，利用θ_z和$\varphi_{b|z}$来确定n_z和$n_{b|z}$。

$$\theta_z = \frac{n_z + \sum_{K=1}^{K}(\gamma * A) + \alpha}{|B| + \sum_{l=1}^{l}(\gamma * A) + K\alpha} \tag{6}$$

$$\varphi_{b|z}=\frac{n_{b|z}+\sum_{K=1}^{K}(\gamma*A)+\beta}{\sum_{b}n_{b|z}+2\sum_{K=1}^{K}(\gamma*A)+M\beta} \tag{7}$$

公式(6)中*l*表示满足条件的所有扩展词对的数量。

3.3 算法描述

根据上节的改进思想，具体的算法流程如下，因为对传统的采样过程添加了一些额外的操作，会影响到采样的平衡性，容易导致最后的主题矩阵值为负，即$P_j<0$，此时，如公式(8)所示，直接将主题矩阵值取正，作为最终的主题矩阵值，使吉布斯采样达到一种平衡稳态。

$$P_j=-P_j \tag{8}$$

Input：主题数量K，双词组合数量|B|，迭代次数N，α，β，C，P
Output：主题分类结果
利于cw2vec模型计算语义距离：
    for B=1 to |B|
        for $b_i,b_j\in$ B do
            按照公式(1)计算语义相似度
所有词对主题初始化：
    for B=1 to |B|
        for $b_i,b_j\in$ B do
            if γ>C
                公式(2)
            else
                公式(3)
吉布斯采样过程：
    for i=1 to N do
        for $b_i,b_j\in$ B do
            if γ>C
                按照公式(4)计算出每次更新的主题
        for j in P：
            if $P_j<0$：
                按照公式(8)转换矩阵值
                更新吉布斯采样的各个参数，n_z，$n_{b|z}$和$n_{b'|z}$
            else
                按照公式(5)计算出每次更新的主题
        for j in P：
            if $P_j<0$：
                按照公式(8)转换矩阵值
                更新吉布斯采样的各个参数，n_z，$n_{b|z}$和$n_{b'|z}$
分别按照公式(6)和公式(7)计算出θ_z和$\varphi_{b|z}$

4 实验及结果分析(Experiments and results analysis)

4.1 实验数据及参数设置

实验数据主要是使用基于python的PySpider框架爬取各大电商网站的冰箱评价，集中于某一类商品评价是为了在测试主题模型的时候，能够得到主题的大致范围，便于分析。实验中我们采集了共500000条评论，采用十折交叉验证法来处理语料。经过去停用词、分词等预处理操作之后的部分实验数据如图4所示。

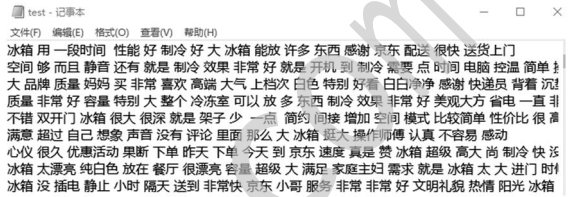

图4 预处理后部分实验数据

Fig.4 Some experimental data after pretreatment

为了能充分证明融合cw2vec模型给主题模型带来的积极效果，本文同时使用了word2vec进行了对比实验，为了适应短文本，实验参数设置如表4所示。

表4 实验模型参数设置表

Tab.4 Experimental model parameter setting table

参数名	参数值
词向量维度	200
窗口大小	10
负采样数目	15
迭代次数	20
最小词频	5
学习率	word2vec(0.025) cw2vec(0.025)
n-gram特征	minn=3，maxn=18

4.2 对照实验及结果分析

(1)实验测评标准

在评估主题模型时，大多采用主题凝聚度和*KL*散度这两个指标。前者主要反映主题的内聚程度，而后者反映主题的差异性。本文具体所用指标如下：

①主题凝聚度

*TC*值越高，说明主题的内聚性就越强，公式如下：

$$TC(t;B^{(t)})=\sum_{m=2}^{M}\sum_{l=1}^{m-1}\log\frac{N(b_m^{(t)},b_l^{(t)})+1}{N(b_l^{(t)})} \tag{9}$$

其中，$N(b)$表示文档中出现词b的文档的数目；$N(a,b)$表示文档中同时出现词a和b的文档的数目。

②Jensen−Shannon距离

传统的KL距离，用来评价两个概率分布之间的差异大小。KL距离越大，证明获得的主题质量越高。KL散度公式如下：

$$D_{KL}(p\|q)=\sum_i p(i)\log\frac{p(i)}{q(i)} \tag{10}$$

p和q分别表示不同主题下的主题−词分布，i参数表示主题—词分布的数量。

因此这里采用Jensen—Shannon距离，JS距离的定义为：

$$Js=\frac{1}{2}KL(p\|\frac{p+q}{2})+\frac{1}{2}KL(q\|\frac{p+q}{2}) \tag{11}$$

Js表示根据平均距离算出的KL距离。

(2)定性评估

首先，随机抽取冰箱的几个属性词，然后通过cw2vec模型训练这几个属性词，确定余弦距离值最大的四个词语，作为相关词。部分结果展示如表5所示。

表5 模型训练得到的语义相关度

Tab.5 Semantic relevance obtained from model training

中心词	相关词	余弦距离	中心词	相关词	余弦距离
冰箱	电器	0.72341	价格	便宜	0.68418
	家电	0.71884		划算	0.66920
	品牌	0.68432		实惠	0.65341
	宝贝	0.64239		抢购	0.58723
快递	物流	0.80973	外观	漂亮	0.70145
	送货	0.67216		空间	0.54028
	发货	0.61734		大气	0.53926
	配送	0.56178		时尚	0.47163

由上表可以看出，与中心词“冰箱”语义相关度较高的有“电器”“家电”；而与“价格”语义相关较高的词语有“便宜”“划算”和“实惠”，表中可看出三者的余弦距离值差异很小，符合人们日常用语习惯。证明模型达到实际期望。

(3)定量评估

①不同距离阈值C对词对采样数量的对比

由本文前面的内容可知，阈值C决定了词对的扩展数量。图5表示在吉布斯采样过程中，词对采样数量随着阈值C的增长而呈下降趋势的情况，且C=0时，语料中词对的数量在$1.4*10^6$左右，C=0.1时，数量增加到$3.8*10^6$，证明基于语义阈值C的词对数量扩展方式的有效性。

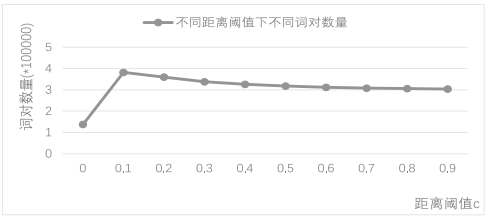

图5 距离阈值C与词对采样数量的关系

Fig.5 Relationship between the distance threshold C and the number of word pair samples

②不同距离阈值C的对比

由本文内容分析可知，最后主题的凝聚度不仅取决于改进模型的词向量训练过程，还受到语义阈值C的影响。那么到底阈值C取多少最为合适，我们通过取不同的阈值C来对改进的主题模型进行实验，取最好的TC值所对应的阈值C为最佳阈值。实验结果如图6所示，在主题数量为5时，不同语义阈值C所对应的TC值最稳定。从图中折线的走势得出，在语义距离阈值C接近0.4时，所有TC值均呈上升趋势，总体取得的效果最好。

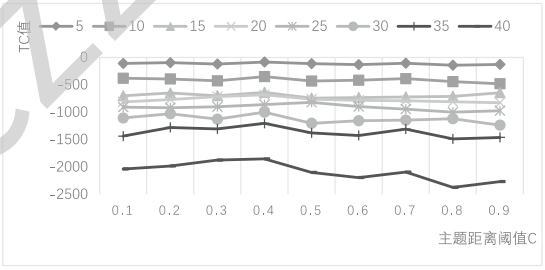

图6 不同距离阈值C的不同TC值

Fig.6 Different TC values for different distance thresholds C

③主题凝聚度

为了验证改进的主题模型的有效性，我们将其与其他主题模型进行对比实验，实验中，先验参数设置为$\alpha=50/K$，$\beta=0.01$，由上节实验结果得出距离阈值C=0.4。以上三个参数确定后，我们通过评测标准主题凝聚度TC来验证主题模型的有效性。实验时，我们将抽取主题数量定为5、10、15、20、25、30、35、40。对比实验结果如图7所示。

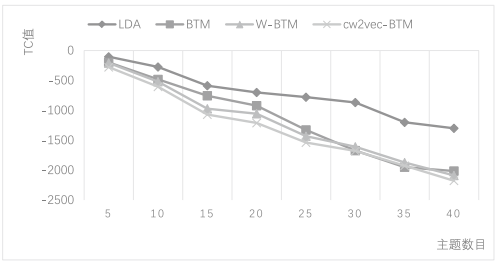

图7 主题凝聚度对比

Fig.7 Comparison of subject cohesion

可以看出，主题数目为5、10时，改进模型与传统模型的*TC*值几乎相同，LDA模型的*TC*值略低于BTM模型及其改进模型，证明LDA模型不适合处理短文本，主题聚类效果差，所有模型的*TC*值都随着主题数量的增加而增长，且不同模型之间的效果区分也越来越明显；很明显，融入了词向量的改进模型的*TC*值都有所提升，其中结合cw2vec-BTM模型的改进模型的*TC*值是最高的，主题凝聚效果最好。

④Jensen-Shannon距离

*JS*距离越大，主题质量越好。

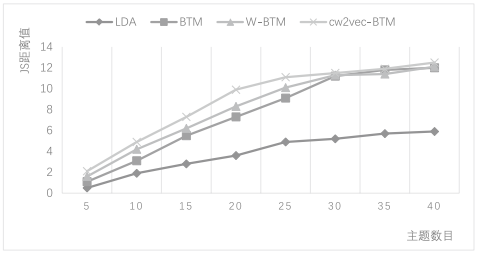

图8 不同模型的JS距离对比

Fig.8 Comparison of JS distances of different models

由图8可以看出，BTM及改进模型的效果要优于LDA模型，比较BTM模型和改进模型可以看出，改进模型在主题数为30之前，*JS*距离值高于BTM模型，但随着主题数目的增加，*JS*距离值被BTM反超。出现这种情况可能是由于随着主题数目的增加，不断扩展双词的数量，导致每个主题中语义相关的词语越来越多，最终每个主题的内聚性越来越强，但主题之间的差异变得越来越不明显；比较传统模型及其改进模型，融入了词向量后模型的*JS*距离明显增大，其中cw2vec-BTM模型的效果最优。

5 结论(Conclusion)

针对BTM主题模型建模过程中，语料库中的词对之间没有相互的语义联系的问题，本文提出了一种改进的BTM主题模型算法，在BTM主题模型的基础上，借助深度学习的cw2vec模型来训练词向量，给共现的双词融入更加精准的语义关系。最后，与传统的LDA和BTM模型、融入word2vec模型BTM模型进行对比，取得了最优的主题聚类效果，证明了本文所提方法的通用性与有效性。

参考文献(References)

[1] Zhu L,Wang G,Zou X.A Study of Chinese Document Representation and Classification with Word2vec[C].2016 9th International Symposium on Computational Intelligence and Design(ISCID).IEEE,2016.

[2] Ali M,Khalid S,Aslam M H.Pattern Based Comprehensive Urdu Stemmer and Short Text Classification[J].IEEE Access,2017(99):1.

[3] Li P,He L,Wang H,et al.Learning From Short Text Streams With Topic Drifts[J].Cybernetics IEEE Transactions on,2018,48(9):2697-2711.

[4] Blei D M,Ng A Y,Jordan M I,et al.Latent Dirichlet Allocation[J].Journal of Machine Learning Research,2003(3):993-1022.

[5] L.Hong and B.Davison.Empirical study of topic modeling in Twitter,in Proceedings of the First Workshop on Social Media Analytics.ACM,2010:80-88.

[6] Yan X,Guo J,Lan Y,et al.A biterm topic model for short texts[C].Proceedings of the 22nd international conference on World Wide Web.ACM,2013:1445-1456.

[7] Cheng Z,Wenxiu W U,Ning D.Improved short text classification method based on BTM topic features[J].Computer Engineering and Applications,2016,13(52):95-100.

[8] 张芸.基于BTM主题模型特征扩展的短文本相似度计算[D].安徽大学,2014.

[9] Cao S,Lu W,Zhou J,et al.cw2vec:Learning chinese word embeddings with stroke n-gram information[C].Lousiana:Thirty-Second AAAI Conference on Artificial Intelligence,2018:1-8.

[10] Mikolov T,Chen K,Corrado G,et al.Efficient Estimation of Word Representations in Vector Space[Online].available:http://arxiv.org/abs/1301.3781,September 7,2013.

[11] Chen X,Xu L,Liu Z,et al.Joint learning of character and word embeddings[C].International Conference on Artificial Intelligence.AAAI Press,2015.

[12] Su T R,Lee H Y.Learning Chinese Word Representations From Glyphs Of Characters[C].Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,2017:264-273.

[13] Yu J,Jian X,Xin H,et al.Joint embeddings of chinese words,characters,and fine-grained subcharacter components[C].Proceedings of the,2017 Conference on Empirical Methods in Natural Language Processing.2017:286-291.

作者简介：

王云云(1992-)，女，硕士生.研究领域：软件工程技术.

张云华(1965-)，男，博士，研究员.研究领域：软件工程.