Journal of Computer Applications
计算机应用,2016,36(10):2794-2798
ISSN 1001-9081
CODEN JYIIDU
2016-10-10
http://www.joca.cn

文章编号:1001-9081(2016)10-2794-05
DOI:10.11772/j.issn.1001-9081.2016.10.2794

基于表示学习的中文分词

刘春丽,李晓戈*,刘 睿,范 贤,杜丽萍
(西安邮电大学 计算机学院,西安 710121)
(* 通信作者电子邮箱 lixg@xupt.edu.cn)

摘 要: 为提高中文分词的准确率和未登录词(OOV)识别率,提出了一种基于字表示学习方法的中文分词系统。首先使用 Skip-gram 模型将文本中的词映射为高维向量空间中的向量;其次用 *K*-means 聚类算法将词向量聚类,并将聚类结果作为条件随机场(CRF)模型的特征进行训练;最后基于该语言模型进行分词和未登录词识别。对词向量的维数、聚类数及不同聚类算法对分词的影响进行了分析。基于第四届自然语言处理与中文计算会议(NLPCC2015)提供的微博评测语料进行测试,实验结果表明,在未利用外部知识的条件下,分词的 *F* 值和 OOV 识别率分别达到 95.67% 和 94.78%,证明了将字的聚类特征加入到条件随机场模型中能有效提高中文短文本的分词性能。

关键词: 表示学习;词向量;聚类;条件随机场;中文分词
中图分类号: TP391.1 **文献标志码:** A

Chinese word segment based on character representation learning

LIU Chunli, LI Xiaoge*, LIU Rui, FAN Xian, DU Liping
(*College of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an Shaanxi* 710121, *China*)

Abstract: In order to improve the accuracy and the Out Of Vocabulary (OOV) recognition rate of the Chinese word segmentation, a Chinese word segmentation system based on character representation learning method was proposed. Firstly, the word in the text was mapped to a vector in a high-dimentioanl vecter space using Skip-gram model; then the *K*-means clustering algorithm was used to acquire clusters of the word vector, and the clustering results were regarded as features of Conditional Random Fields (CRF) model for training. Finally the CRF model was used for word segmentation and OOV recognition. The influences of the word vector dimensions, the number of clusters and different cluster algorithm on word segmentation were analyzed. Experiments were conducted on the 4th CCF Conference on Natural Language Processing & Chinese Computing (NLPCC2015) corpus. Experimental results show that the proposed system can effectively improve Chinese short text segmentation performance without using external knowledge, the *F*-value and the OOV recognition rate achieve to 95.67% and 94.78% respectively.

Key words: representation learning; word vector; clustering; Conditional Random Field (CRF); Chinese word segmentation

0 引言

词是能够独立运用的最小语言单元。英文中,单词之间以空格作为自然分界符,而中文则是以字为基本的书写单位,词语之间没有明显的分界符,如果不进行分词,那么计算机就无法得知中文词语的确切边界,因而很难理解文本中所包含的语义信息。因此,中文分词是中文自然语言处理的一项基础性工作,其在信息检索、文本自动分类及数据挖掘等领域具有举足轻重的地位,是中文信息处理技术发展的技术瓶颈,研究中文分词具有十分重要的意义。

中文分词和未登录词(Out Of Vocabulary, OOV)识别都可以看作一个序列标记任务来完成,应用广泛的序列标注模型主要有隐马尔可夫模型(Hidden Markov Model, HMM)[1]、最大熵马尔可夫模型(Maximum Entropy Markov Model, MEMM)[2]和条件随机场(Conditional Random Field, CRF)[3]模型等。相对于 HMM 和 MEMM,CRF 模型能灵活地选择特征和控制训练数据的拟合程度。基于 CRF 模型进行序列标记的主要任务是针对特定的问题选择有效的特征来描述特定的语言内容。传统的 CRF 模型中所用的特征表示大多都是基于词袋模型的,而这种模型会出现"词汇鸿沟"现象,即任意两个词之间都是孤立的,即使是"话筒"和"麦克风"这样的同义词也不能幸免于难。同时,使用词袋作为特征训练模型时,低频词容易造成训练不足,也有可能出现过拟合。

2006 年 Hinton 等[4]提出深度学习后,从大量未标记文本中学习字表示的方法已经被证实对识别未登录词、命名实体识别、词性标注[5]和依存分析[6]等是有效的。比如:Mann

收稿日期:2016-03-24;修回日期:2016-06-21。 基金项目:国家自然科学基金资助项目(61373116);陕西省普通高等学校重点学科专项资金资助项目(112-1602);西安邮电大学研究生创新基金资助项目(ZL2013-30)。

作者简介:刘春丽(1990—),女,山西临汾人,硕士研究生,主要研究方向:自然语言处理、文本数据挖掘; 李晓戈(1962—),男,安徽合肥人,教授,博士,主要研究方向:自然语言处理、机器学习、数据挖掘; 刘睿(1992—),男,陕西咸阳人,硕士研究生,主要研究方向:自然语言处理、大数据; 范贤(1991—),女,陕西咸阳人,硕士研究生,主要研究方向:情感分析、大数据; 杜丽萍(1987—),女,陕西宝鸡人,硕士研究生,主要研究方向:自然语言处理、大数据。

等[7]提出一个半监督 CRF 的方法来提高序列分词和词性标注的准确率; Yu 等[8]提出一个深层结构的 CRF 序列标注模型; Zheng 等[9]利用深度学习的方法进行中文分词和词性标注的任务,取得的最好结果的 F 值和 OOV 召回率分别为 95.23%和 72.38%; 来斯惟等[10]将从大规模中文语料中学习得到的语义向量应用到有监督的中文分词中,最终取得 94.90%的 F 值和 81.5%的 OOV 召回率,证明了字向量加入的有效性。这种字表示方法是通过训练数据中的词来学习词向量或词的聚类特征等字表示,然后通过这些字符表示特征总结出词的特征,再在这种特征的基础上进一步进行模型的学习,这样可以显著地提高分词的性能。

本文提出了一个结合字表示和条件随机场的简单有效的半监督方法来提高中文分词的准确率和召回率。首先从大量未标记的微博语料中学习中文字符的语义向量,基于这些语义向量作 K-means 聚类,同时对中文字符进行布朗聚类,然后再将这些字表示特征应用到 CRF 模型中进行有监督的中文分词。与传统的分词方法相比,该方法着重从未标记文本中挖掘逐点互信息和访问多种特征。以 NLPCC2015 中的微博分词任务提供的测试集[11]作测试,本文取得的最好结果为 95.67%的 F 值和 94.78%的 OOV 召回率。实验结果表明字符的表示学习方法对提高中文分词的准确率和召回率是有效的。

1　字表示学习

1.1　系统框架

图 1 描述了系统的整体框架。首先,用空格将原始文本(没有分隔符)的每个字分开,如给出一个句子 $S=[C_0C_1C_2C_3C_4]$,每个字为 C_i,将其分割为字块 $[C_0,C_1,C_2,C_3,C_4]$; 其次,利用 Google 开源的深度学习工具 word2vec[12]对预处理后的语料进行学习得到字符的向量表示; 再次,利用 K-means 聚类算法得到字的一种聚类类别,同时,使用布朗聚类算法得到字的另一种聚类类别; 最后,将这两种不同的聚类结果作为 CRF 的特征训练语言模型。

本文将 K-means 聚类类别和布朗聚类类别同时作为特征加入到 CRF 模型中,可以降低某些不可靠特征造成的风险,也自然地解决了在缺乏词语分隔符的原始文本中进行表示学习的问题。

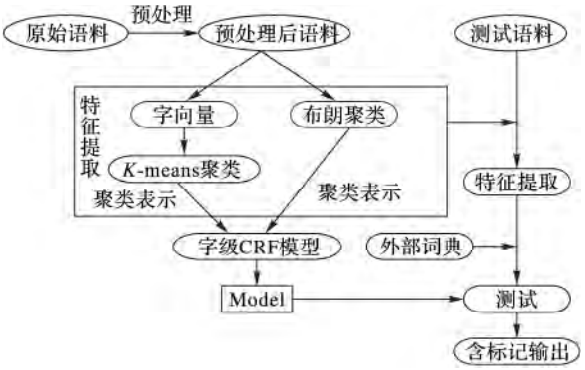

图 1　系统框架

1.2　基于向量的字表示

基于向量的字符表示学习是将一个字用一个低维实向量来表示,向量中的每一维表示某种隐含的句法或语义信息,目标是把对文本内容的处理简化为向量空间中的向量运算,计算出字符在向量空间上的相似度,以此来表示文本语义上的相似度。本文使用 Google 基于深度学习的开源工具 word2vec 生成词向量。用 $\boldsymbol{v}_c \in \mathbf{R}^d$ 表示本文想要学习的字符的 d 维向量; ***Context*** 表示当前字符的上下文,即当前字符前面和后面各若干个字符(一般为 1 ~ 5 的一个随机数)。根据神经语言模型[13]的研究,可以通过最大化以下目标来优化并得到 $\boldsymbol{v}_c$:

$$\sum_c \log p(\boldsymbol{v}_c \mid \boldsymbol{Context}) = \sum_c \log\left[\frac{\exp(\boldsymbol{v}_c^{\mathrm{T}} \cdot \boldsymbol{Context})}{\sum_w \exp(\boldsymbol{v}_w^{\mathrm{T}} \cdot \boldsymbol{Context})}\right] \quad (1)$$

由式(1)可知,通过对目标概率模型 $p(\boldsymbol{v}_c \mid \boldsymbol{Context})$ 使用极大似然估计法,可求得最优化的 $\boldsymbol{v}_c$,即本文想要学习的字符向量。而 word2vec 认为 $p(\boldsymbol{v}_c \mid \boldsymbol{Context})$ 这个条件概率可以用能量模型来表示,且定义了一个三非常简单的能量函数,即

$$E(\boldsymbol{v}_c, \boldsymbol{Context}) = -(\boldsymbol{v}_c \cdot \boldsymbol{Context})$$

其中: "·" 表示两个向量的点积。为了计算条件概率 $p(\boldsymbol{v}_c \mid \boldsymbol{Context})$,需要把语料库里面所有的字符的能量都计算一次,式(1)中的归一化分母 $\sum_w \exp(\boldsymbol{v}_w^{\mathrm{T}} \cdot \boldsymbol{Context})$ 即表示所有字符(如词汇量)的和。由此,再根据 $\boldsymbol{v}_c$ 的能量,便可得到一个比值,这个比值即为出现 $\boldsymbol{v}_c$ 条件概率。式(1)的目标是在优化概率模型的过程中为大量未标记文本生成了所有字符的向量表示。根据文献[5],为了保证各个字符向量之间具有可比性,对这些字符向量进行了归一化。

根据式(1)学习得到的词向量特征是稀疏的、离散的,虽然离散的特征可能更易集成到现有的自然语言处理(Natural Language Processing,NLP)系统中去,但却丧失了对实向量辨别相似之处的能力,所以对这些实值向量进一步作 K-means 聚类。在每次迭代过程中,经典的 K-means 方法将数据对象归入相距中心点最近的一类,同时重新调整和计算这些类的中心点,直到中心点收敛于确定的位置。K-means 算法计算简单、有效,但也存在一些严重的不足,需要事先确定聚类的数目,然而,在一般情况下,这个数目往往是不能准确得到的; 同时,聚类的结果和效率也往往会受到初始聚类中心位置的影响,在数据维数较高时,聚类的质量也明显下降。本文使用的这种两步的聚类方法其中心向量是从 word2vec 中得出的,并且实验结果验证了这种方法是有效的[14]。

1.3　基于聚类的字表示

本文使用的基于聚类的表示方法是被广泛使用的布朗聚类算法,利用它来训练大量未标记的原始文本得到字的聚类特征。布朗聚类是一个分级聚类算法,这个算法通过最大化二元互信息对字组进行聚类,所以布朗聚类是一个基于类别的两元语言模型。其运行的时间复杂度是 $O(V \cdot K^2)$,其中 V 是词汇的大小,K 是聚类的个数。聚类的分级特性意味着可以在层次结构中选择若干个级别的词类别,这样可以弥补一些由少量字组构成的稀少类别。布朗聚类的一个缺点是其完全基于两字组统计数据,而并没有考虑更大的上下文字组的用法。以前的工作显示词聚类对一般的命名实体识别是一个好

的特征[15]，Turian等[5]用于基于布朗聚类的字表示来提升命名实体识别在新闻领域的识别。本文采用布朗聚类，从一个分层型字聚类算法中创建聚类特征，给每个字分配一个基于哈夫曼编码的二进制表示，如“秒”的二进制表示为“0101100”，将这个二进制数作为这个字的聚类特征。

本文使用的两种不同的聚类表示方法的区别在于对字的表示方法不同，*K*-means聚类算法是以字向量表示为基础的，而布朗聚类则是基于字本身特征的二进制编码表示，在2.3节的实验结果分析中证明了这两种不同的字聚类表示特征对分词结果产生了不同程度的影响。

2 实验及结果分析

2.1 数据和初始化

采用NLPCC2015中文微博文本分词任务提供的训练及测试语料作为本文的训练及测试语料，其中训练语料1.96 MB(共10 000个句子，215 567个词)，测试语料662 KB(共5 000个句子，106 843个词)。

本文使用条件随机场开源工具CRF++[16]完成中文微博数据上的分词处理。与其他分词方法类似，本文采用四词位标注集体系对汉字进行标注，即对于多字词，词首汉字标签为B，词尾汉字标签为E，词中汉字标签为M；对于单字词，其标签为S。有研究统计表明，在所有的语料中，90%的词是由1～2个汉字组成，95%的词是由3个或3个以下的字构成，99%的词是由5个或5个以下的字构成[17]，也就是说中文词语的长度绝大多数在5字以内，所以本文采用10特征模板集，即以当前汉字为基础取其前后各两个汉字作为上下文，以5字长度的上下文窗口来作特征信息统计。设当前字符为C_i，上下文为$\cdots C_{i-1}C_iC_{i+1}\cdots$，baseline特征实例[18]如下：

单字字符特征：$C_s(i-3<s<i+3)$；

双字字符特征：$C_sC_{s+1}(i-3<s<i+2)$；

字符双连词特征：$C_sC_{s+1}(i-2<s<i+2)$；

字符相同跳跃特征：$C_s=C_{s+1}(i-4<s<i+2)$。

本文将字表示特征加入到CRF模型中，例如，对于目标字符C_s，分别提取基于字符的向量表示、*K*-means和布朗聚类训练得到特征并加入到CRF模型中。实验中字符的向量表示是用word2vec训练得到的字符向量$\boldsymbol{R}^d$。在这些字符向量的基础上利用*K*-means聚类，经过一组对比实验(如图2所示)可以得到，当向量维度$d=200$，*K*-means聚类类别数$k=400$时分词效果较好，故在后续特征叠加的实验中均以此二值得到的结果进行叠加。对布朗聚类，本文采用聚类所得的二进制编码作为特征加入CRF模型。

2.2 评价标准

本文选用的评测指标为准确率(P)、召回率(R)和综合评价指标F值。具体定义如下：

$$P=\frac{\text{系统正确识别的词语总数}}{\text{系统识别的词语总数}}\times 100\%$$

$$R=\frac{\text{系统正确识别的词语总数}}{\text{测试语料中的词语总数}}\times 100\%$$

$$F=\frac{2\times P\times R}{P+R}\times 100\%$$

2.3 实验结果及分析

以基于字符级不加特征的CRF分词作为baseline。word2vec及布朗聚类采用NLPCC2015分词任务中提供的微博训练语料所得，其中含10 000个句子，共215 567个词。

如图2所示，探讨了利用word2vec训练得到的不同维度的字向量以及*K*-means聚类类别数不同的情况下对分词结果的影响，图中以分词结果F值作为参考。从图2中曲线可看出，当向量维度$d=200$，*K*-means聚类类别数$k=400$时分词效果较好，F值达到94.07%。同时，随着字向量维度d值的增大，分词结果F值显得更平滑，当$d=50$时，分词结果F值显得最不稳定，震荡较严重；而当$d=500$时，分词结果F值的折线图近似为一条平滑的曲线，这意味着随着向量维度的增大，向量所表示的字得语义信息更为全面和准确，聚类特征表现得更为良好，所以分词结果表现更趋于平滑。

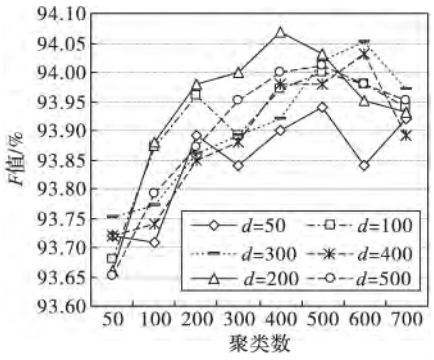

图2 字符向量维度及*K*-means聚类个数

从聚类结果中也可看出，当字向量维度增大时，聚类结果中各类别所包含的字的语义表现得更为相近。如表1所示，列出了在字向量维度分别为50和200时得到的与“海”“刘”最相近的字。可以看出，当字向量维度$d=200$时，类别内部的字之间的相似度更近，语义更相关，而当字向量维度$d=50$时，结果没有$d=200$时的表现好。比如当$d=50$时，与“海”字语义相近程度排在前十的字中出现了“莞”字。

表1 字向量维度不同时得到的与“海”“刘”最相近的字

字	字向量维度	相似的字
海	50	南 琼 廊 岸 屯 湖 潭 河 莞 丘
	200	琼 岸 潭 滩 湖 岛 河 岭 云 峡
刘	50	彭 徐 杰 蔡 李 魏 赵 锡 柯 林
	200	蔡 彭 徐 魏 宋 蒋 赵 杰 吕 唐

图3对比了布朗聚类与字向量分别为200维和300维时*K*-means聚类在类别数不同时对分词结果的影响。由图3可以看出，当类别数$k=200$时，布朗聚类对分词结果的影响达到峰值，此后分词效果随着k值的增加而递减。整体而言，当类别数k较小时，布朗聚类效果比*K*-means聚类效果更好，随着k值的增大，布朗聚类的效果要比*K*-means聚类的效果下降更快。

综上可知，仅加入字向量的*K*-means聚类特征时，在$d=200$，$k=400$时分词效果达到最优；仅加入字的布朗聚类特征时，在$k=200$时CRF分词效果达到最优。所以本文以这两组最优的数据作参考，将这两种特征结合起来以达到提高CRF分词精度的目的。表2中所列*K*-means聚类以$d=200$，

k = 400时数据作参考,布朗聚类以 k = 200 时数据作参考。表 2中,w2v 为加入的利用 word2vec 及 K-means 所得特征,Brown 为布朗聚类所得特征。相对 baseline 结果而言,两种不同的字表示方法的加入均对分词结果有积极的作用,且布朗聚类更优于 K-means 聚类。同时也可看出,当两者叠加使用时分词效果比任意只使用其中一种时的效果要好,此时 F 值达到 94.44%,OOV 召回率达到92.91%,相对 baseline 分别提高了 1.28 个百分点和 1.41 个百分点,这说明这两种不同的聚类表示提供了不同的信息,弥补了各自部分的缺点。

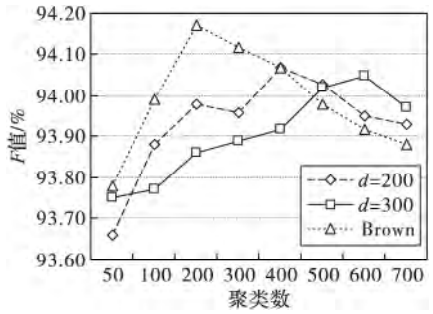

图 3　布朗聚类与 K-means 聚类对分词结果影响的对比

表 2　特征叠加实验结果　　%

实验方案	准确率	召回率	F 值	OOV 召回率
baseline	93.39	92.93	93.16	91.50
baseline + w2v	94.24	93.89	94.07	92.51
baseline + Brown	94.33	94.01	94.16	92.72
baseline + w2v + Brown	94.57	94.30	94.44	92.91
baseline + Dic	94.96	95.34	95.15	94.04
baseline + w2v + Brown + Dic	95.48	95.86	95.67	94.78

由于 OOV 词影响分词精度的主要因素[19],本文通过引入词典知识(表 2 中的 Dic)来降低未登录词和交叠歧义对分词结果的影响,词典中的词均来源于 NLPCC 提供的已有正确标记序列的微博语料。测试结果显示,引用词典特征后,F 值和 OOV 召回率比 baseline 的结果分别提高了 1.99 个百分点和 2.54 个百分点,说明了词典引入的重要性。在引入词典特征的基础上叠加使用两种字表示特征后,F 值和 OOV 召回率比仅加入词典特征的结果分别提高了 0.52 个百分点和 0.74 个百分点,这也体现了字表示对中文分词结果起改善作用。

在 NLPCC2015 分词任务中 Closed track 的部分中 Min[20] 等使用线性链条件随机场加规则过滤的方法取得了95.03%、95.03%、95.03%的准确率、召回率和 F 值,本文结果分别比其提高了 0.45 个百分点、0.83 个百分点和 0.64 个百分点,说明本文方法可以有效改进中文分词。

3　结语

本文探索了一种简单有效的分词方法,该方法分别使用字符的语义向量、K-means 聚类和布朗聚类得到不同的字符表示特征,再将这些字符表示特征应用到 CRF 模型中做训练,以 NLPCC2015 分词任务提供的微博测评语料进行测试,最好的结果准确率、召回率、F 值及 OOV 识别率分别达到 95.48%、95.86%、95.67%和 94.78%。实验结果表明,这种将字符表示特征加入到 CRF 模型中的半监督方法对改善中文分词的结果是有效的,但是目前这种方法仍然无法完全取代人工设计特征的有监督的学习方法。未来的工作将从以下两个方面尝试: 1) 对语料中不同长度的字块进行表示学习,比如 2-gram 字块和 3-gram 字块,将其加入到 CRF 模型中,通过多长度的表示学习来提高分词准确率; 2) 增加外部知识学习,扩大语料库,进行开放测试并提升其分词效果。

参考文献:

[1] 魏晓宁. 基于隐马尔可夫模型的中文分词研究[J]. 电脑知识与技术(学术交流), 2007, 4(11): 885 - 886. (WEI X N. HMM-based of study on Chinese language classifying words [J]. Computer Knowledge and Technology (Academic Exchange), 2007, 4(11): 885 - 886.)

[2] ANDREW M, DAYNE F, FEMANDO P. Maximum entropy Markov models for information extraction and segmentation [C] // Proceedings of the Seventeenth International Conference on Machine Learning. New York: ACM, 2000: 591 - 598.

[3] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C] // Proceedings of the 18th International Conference on Machine Learning. New York: ACM, 2001: 282 - 289.

[4] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504 - 507.

[5] TURIAN J, RATINOV L, BENGIO Y. Word representations: a simple and general method for semi-supervised learning [C] // Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010: 384 - 394.

[6] KOO T, CARRERAS X, COLLINS M. Simple semi-supervised dependency parsing [C] // Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2008: 595 - 603.

[7] MANN G S, MCCALLUM A. Generalized expectation criteria for semi-supervised learning of conditional random fields [C] // Proceedings of the 2008 Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010: 1374 - 1377.

[8] YU D, WANG S, DENG L. Sequential labeling using deep-structured conditional random fields [J]. IEEE Journal of Selected Topics in Signal Processing, 2010, 4(6): 965 - 973.

[9] ZHENG X Q, CHEN H Y, XU T Y. Deep learning for Chinese word segmentation and POS tagging [C] // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: [s. n.], 2013: 647 - 657.

[10] 来斯惟, 徐立恒, 陈玉博, 等. 基于表示学习的中文分词算法探索[J]. 中文信息学报, 2013, 27(5): 8 - 14. (LAI S W, XU L H, CHEN Y B, et al. Chinese word segment based on character representation learning [J]. Journal of Chinese Information Processing, 2013, 27(5): 8 - 14.)

[11] QIU X, QIAN P, YIN L, et al. Overview of the NLPCC 2015 shared task: Chinese word segmentation and POS tagging for micro-blog texts (2015) [EB/OL]. [2015-03-10]. http://arxiv.org/abs/1505.0759.

[12] word2vec [EB/OL]. [2015-03-12]. https://code.google.com/p/word2vec/.

[13] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [EB/OL]. [2015-03-10]. https://arxiv.org/abs/1310.4546.

[14] WU X, ZHOU J, SUN Y et al. Generalization of words for Chinese dependency parsing [C]// Proceedings of the 4th CCF Conference on Natural Language Processing and Chinese Computing, LNCS 9362. Berlin: Springer, 2015: 36-46.

[15] MILLER S, GUINNESS J, ZAMANIAN A. Name tagging with word clusters and discriminative training [EB/OL]. [2015-03-10]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.105.9395.

[16] CRF++ [EB/OL]. [2015-03-20]. http://sourceforge.net/projects/crfpp/.

[17] GAO J F, LI M, WU A, et al. Chinese word segmentation and named entity recognition: a pragmatic approach [J]. Computational Linguistics, 2005, 31(4): 531-574.

[18] SUN X, WANG H, LI W. Fast online training with frequency-adaptive learning rates for Chinese word segmentation and new word detection [C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2012: 253-262.

[19] 杜丽萍, 李晓戈, 于根, 等. 基于互信息改进算法的新词发现对中文分词系统改进 [J]. 北京大学学报(自然科学版), 2016, 52(1): 35-40. (DU L P, LI X G, YU G, et al. New word detection based on an improved PMI algorithm for enhancing segmentation system [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2016, 52(1): 35-40.)

[20] MIN K R, MA C G, ZHAO T M, et al. BonsonNLP: an ensemble approach for word segmentation and POS tagging [C]// Proceedings of the 4th CCF Conference on Natural Language Processing & Chinese Computing. Berlin: Springer, 2015: 520-526.

Background

This work is partially supported by the National Natural Science Foundation of China (61373116), the Development Funds for the Key Subjects of the Universities in Shaanxi Province (112-1602), the Graduate Innovative Foundation of Xi'an University of Posts & Telecommunications (ZL2013-30).

LIU Chunli, born in 1990, M. S. candidate. Her research interests include natural language processing, text data mining.

LI Xiaoge, born in 1962, Ph. D., professor. His research interests include natural language processing, machine learning, data mining.

LIU Rui, born in 1992, M. S. candidate. His research interests include natural language processing, big data.

FAN Xian, born in 1991, M. S. candidate. Her research interests include sentiment analysis, big data.

DU Liping, born in 1987, M. S. candidate. Her research interests include natural language processing, big data.

(上接第2788页)

[12] 冷亚军, 陆青, 梁昌勇. 基于结构相似性的协同过滤推荐算法 [J]. 小型微型计算机系统, 2015, 36(10): 2266-2269. (LENG Y J, LU Q, LIANG C Y. Collaborative filtering recommendation algorithm based on structure similarity [J]. Journal of Chinese Computer Systems, 2015, 36(10): 2266-2269.)

[13] CHOI K, SUH Y. A new similarity function for selecting neighbors for each target item in collaborative filtering [J]. Knowledge-Based Systems, 2013, 37(1): 146-153.

[14] 朱锐, 王怀民, 冯大为. 基于偏好推荐的可信服务选择 [J]. 软件学报, 2011, 22(5): 852-864. (ZHU R, WANG H M, FENG D W. Trustworthy services selection based on preference recommendation [J]. Journal of Software, 2011, 22(5): 852-864.)

[15] 王磊, 赵庆建, 罗兴峰. 基于项目相关度的STI新群体冷启动推荐方法 [J]. 小型微型计算机系统, 2015, 36(3): 450-453. (WANG L, ZHAO Q J, LUO X F. Degree of item correlation based STI for new community cold start recommendation [J]. Journal of Chinese Computer Systems, 2015, 36(3): 450-453.)

[16] BOBADILLA J, ORTEGA F, HERNANDO A. A collaborative filtering similarity measure based on singularities [J]. Information Processing and Management, 2011, 48(2): 204-217.

[17] 夏小伍, 王卫平. 基于信任模型的协同过滤推荐算法 [J]. 计算机工程, 2011, 37(21): 26-28. (XIA X W, WANG W P. Collaborative filtering recommendation algorithm based on trust model [J]. Computer Engineering, 2011, 37(21): 26-28.)

[18] 王茜, 王锦华. 结合信任机制和用户偏好的协同过滤推荐算法 [J]. 计算机工程与应用, 2015, 51(10): 261-270. (WANG Q, WANG J H. Collaborative filtering algorithm combining trust mechanism with user preference [J]. Computer Engineering and Applications, 2015, 51(10): 261-270.)

[19] 张珺, 刘靖, 叶新铭, 等. 基于CPN的可信路由器发现协议建模与仿真分析 [J]. 系统仿真学报, 2012, 24(1): 1-7. (ZHAGN J, LIU J, YE X M, et al. Modeling and simulation of trusted router discovery protocol using colored Petri nets [J]. Journal of System Simulation, 2012, 24(1): 1-7.)

[20] 于洪, 李俊华. 一种解决新项目冷启动问题的推荐算法 [J]. 软件学报, 2015, 26(6): 1396-1406. (YU H, LI J H. Algorithm to solve the cold-problem in new item recommendation [J]. Journal of Software, 2015, 26(6): 1396-1406.)

Background

This work is partially supported by the National Natural Science Foundation of China (61562086, 61462079, 61363083, 61262088).

ZHENG Jie, born in 1992, M. S. candidate. Her research interests include recommendation system, data mining.

QIAN Yurong, born in 1980, Ph. D., associate professor. Her research interests include grid computing, remote sensing image data processing.

YANG Xingyao, born in 1984, Ph. D. His research interests include recommendation system, grid computing, cloud computing, trusted computing.

HUANG Lan, born in 1988, M. S. candidate. Her research interests include big data, data mining, recommendation system.

MA Wanzhen, born in 1992, M. S. candidate. Her research interests include high performance parallel computing.