doi: 10.12068/j.issn.1005-3026.2020.07.005

一种改进的医疗文本分类模型:LS-GRU

李 强[1], 李瑶坤[2], 夏书月[3], 康 雁[1,4]
1. 东北大学 医学与生物信息工程学院, 辽宁 沈阳 110169; 2. 中国石油天然气管道工程有限公司, 河北 廊坊 065000;
3. 沈阳医学院附属中心医院, 辽宁 沈阳 110024; 4. 深圳技术大学 健康与环境工程学院, 广东 深圳 518118)

摘 要: 为了帮助低年资医生阅读胸部CT影像,并更加精确高效地为临床医生反馈影像报告结果,提出一种改进GRU深度学习框架LS-GRU,用来解决影像报告文本分类问题,即可以根据影像科医生描述,自动反馈给临床医生诊断建议.数据来源于呼吸科影像报告1 168例,选择了两种描述相近的疾病(肺气肿和肺炎)进行分类,其中肺气肿患者报告大约652例,肺炎约516例.分别验证GRU、BiGRU及LSTM等模型,实验结果表明,LS-GRU模型分类更精确,且具有较高的鲁棒性.

关 键 词: 深度学习;医疗文本分类;GRU;慢阻肺;LSTM
中图分类号: TG 335.58 **文献标志码**: A **文章编号**: 1005-3026(2020)07-0938-06

An Improved Medical Text Classification Model: LS-GRU

LI Qiang[1], *LI Yao-kun*[2], *XIA Shu-yue*[3], *KANG Yan*[1,4]
(1. College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110169, China; 2. China Petroleum Pipeline Corporation, Langfang 065000, China; 3. The Central Hospital Affiliated to Shenyang Medical, Shenyang 110024, China; 4. College of Health Science and Environmental Engineering, Shenzhen Technology University, Shenzhen 518118, China. Corresponding author: KANG Yan, E-mail: 869242265@qq.com)

Abstract: In order to help radiologists report the CT image results more accurately and effectively to the clinicians, an improved GRU deep learning framework LS-GRU was proposed to solve the classification of image report text, which can be automatically fed back to clinicians according to radiologists' descriptions. The data was collected from more than 1 168 cases of respiratory imaging reports. Two diseases (emphysema and pneumonia) with similar descriptions of radiologists were classified. About 652 cases of emphysema and 516 cases of pneumonia were reported. The GRU, BiGRU and LSTM models were validated, respectively. The results show that the LS-GRU model is more accurate and robust.

Key words: deep learning; medical text classification; GRU(gate recurrent unit); emphysema; LSTM(long-short term memory)

近年来,深度学习在自然语言处理中得到了进一步发展和研究.基于深度神经网络模型实现了在计算机视觉、语音识别、情感分析、自然语言处理等众多领域的应用,同时在各个领域实现了接近甚至超越人类的水平.同时,深度神经网络模型在智能医疗领域得到了长足发展(智能问诊、可穿戴医疗设备、智能诊断等).各大研究机构、团体和科研平台频繁发布大规模标注数据集以及提出了Word2Vec等分布式词向量模型,各种改进的深度神经网络模型在智能医疗领域中不断刷新出更好成绩,其中有很多关键因素如改进模型结构、引入注意力机制的方法等,可以为文档和对象的高层次表示的学习提供更多更高维有效的信息.这对医疗文本精确分类提供了可靠的保障和前提.

随着深度神经网络模型的迅速发展,以及深度神经网络模型在自然语言处理中取得的突出成果,两大主流网络:递归神经网络(recurrent neural

收稿日期: 2019-08-10
基金项目: 国家科技部重点技术研发项目(2017YFC0114200); 国家重点研发计划项目(2018YFC1311900).
作者简介: 李 强(1989-),男,山东曲阜人,东北大学博士研究生; 康 雁(1964-),男,辽宁沈阳人,东北大学教授,博士生导师.

network，RNN）和卷积神经网络（convolutional neural network，CNN）已成为自然语言处理任务的两种主要模型，由于 RNN 在自然语言处理中表现出了杰出性能，尤其在阅读理解和关系推理中取得了成功[1]. 因此本文主要研究并对比 RNN 相关网络模型.

现有的文本分类研究工作包括：Vaswani 等提出的文本注意力机制并详细描述其实现方法[2]；文献[3－5] 在注意力机制文本分类上进一步推广；文献[6－8]分别对循环神经网络模型在文本分类中的应用进行了详细的探讨；Huang 和 Turian 等提出改进的词向量模型方法能够提高文本分类精度[9－10]；许飞飞 等提出一种改进的深度神经网络模型用以文本分类[11].

医疗文本包含了大量丰富的医疗信息（诊断报告、影像报告等），是进行疾病预测、个性化信息推荐、临床决策支持等的重要文本资源[12－13]. 深度学习模型的发展成为了更好分析医疗文本的重要工具. 医疗文本分类最终目的是希望从大量非结构化的自由文本中提取重要信息并加以分析利用[14－15]，为医生的诊断和用药决策提供建议.

本文选择采用 RNN 更为先进的改进网络——GRU（gate recurrent unit）网络作为文本处理的基本网络，为了让提取局部特征以及上下文特征能够关注到非局部特征之间的依赖，本文在 GRU 神经网络的前端加入一层 LSTM（long-short term memory）提取文本特征，后端引入自注意力（self-attention）机制定位分类特征，从而建立了 LS－GRU 网络文本分类模型. 与传统的 GRU 及 LSTM 相比，本文方法解决了 GRU 网络提取特征信息不足及定位特征不准确的问题.

1　实验数据和实验方法

实验数据来源于沈阳医学院附属中心医院 2015—2018 年呼吸科影像报告，选择了其中1 168 例患者的影像报告（肺气肿约 652 例，肺炎 516 例）用于模型训练.

如图 1 所示，实验数据来源于影像报告中的“临床诊断、影像描述”，其中临床诊断作为训练神经网络模型的标签数据，影像描述数据作为训练神经网络模型的输入.

据 2018 年卫计委统计报告显示，全国医疗卫生机构总数达 997 434 个，其中：医院 33 009 个，基层医疗卫生机构 943 639 个，专业公共卫生机构 18 034 个. 如表 1 所示，其中基层卫生机构约占全国总机构数的 96%，为解决基层医院影像医生撰写影像诊断报告的能力较低的问题，尤其对大量的基层医院影像科医生，无法准确判断影像中的表现是何种疾病，基于影像描述的自动判断疾病的功能将可以大大提高影像科医生的判断.

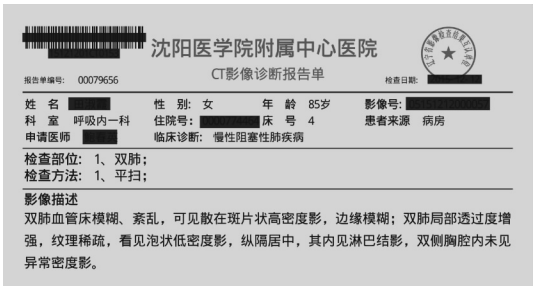
沈阳医学院附属中心医院
CT影像诊断报告单
报告单编号：00079656　　检查日期：[redacted]
姓　名 [redacted]　性　别：女　年　龄 85岁　影像号：[redacted]
科　室 呼吸内一科　住院号：[redacted]　床　号 4　患者来源 病房
申请医师 [redacted]　临床诊断：慢性阻塞性肺疾病
检查部位：1、双肺；
检查方法：1、平扫；
影像描述
双肺血管床模糊、紊乱，可见散在斑片状高密度影，边缘模糊；双肺局部透过度增强，纹理稀疏，看见泡状低密度影，纵隔居中，其内见淋巴结影，双侧胸腔内未见异常密度影。

图 1　慢阻肺影像报告

Fig. 1　An image report of chronic obstructive pulmona disease（COPD）

表 1　全国医疗机构统计表

Table 1　The composition of national medical institutions

名称	机构	比例/%
专业公共卫生机构	妇幼保健机构	0.28
	卫生监督机构	0.27
	疾病预防控制中心	0.32
基层医疗卫生机构	政府办基层医疗卫生机构	11.23
	村卫生室	57.28
	诊所和医务室	21.00
	乡镇卫生院	3.36
	社区卫生服务中心	3.22
医院	未定级医院	0.98
	一级医院	1.00
	二级医院	0.38
	三级医院	0.23

本文提出一种改进的深度神经网络模型用于自动判断影像疾病. 该方法通过对 1 168 例患者影像报告（14 747 个影像句子描述）分析训练，能够很好地对影像描述进行较准确的判断. 训练过程中对数据集按比例 6∶2∶2 分配训练集、验证集和测试集，该过程为随机选取，分配后各数据集中正负样本比例保持不变.

2　模型构建

2.1　主要问题及解决思路

基于影像报告的疾病分类属于一种文本分类问题，其关键是训练基于医疗文本的词向量模型和构建医疗文本分类网络. 基于医疗文本的词向量模型和针对医疗文本分类的深度神经网络模型

可以对医疗文本专业术语进行更好地分类,可以获得更好的分类效果.

本文词向量训练采用 google word2vector 的原理和方法,文本分类模型选择 GRU 网络与 LSTM 网络结合. 具体模型细节如下:

LSTM 网络:长短记忆神经网络通常称作 LSTM,是一种特殊的递归神经网络模型,能够学习长的依赖关系. 由 Hochreiter 提出[7],并被许多人进行了普及和改进. 目前 LSTM 被很好地应用在文本处理中,图 2 为 LSTM 网络结构. LSTM 中有 3 个控制门在 t 时刻输出:遗忘门(f_t),输入门(i_t),输出门(o_t).

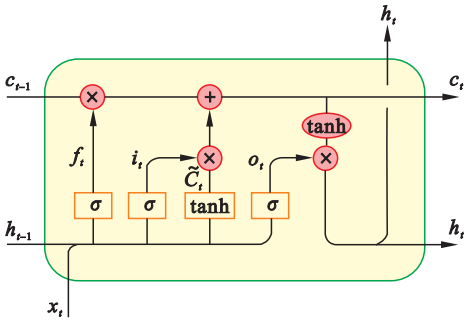

图 2 LSTM 网络结构模型

Fig. 2 LSTM network structure model

GRU 网络是 LSTM 深度网络模型的一个变体,但它只拥有两个门:更新门和重置门,即图 3 中的 z_t 和 r_t. 更新门用于控制上一时刻的输出信息被带入到当前状态中的程度,带出当前时刻状态信息的程度与更新门的值成正比,值越大带出信息程度越大. 重置门作用与更新门相反,带入的信息与值成反比,值越小忽略的程度越大.

因文章重点在模型的设计,因此在 2.2 节中对模型构建过程进行详细介绍.

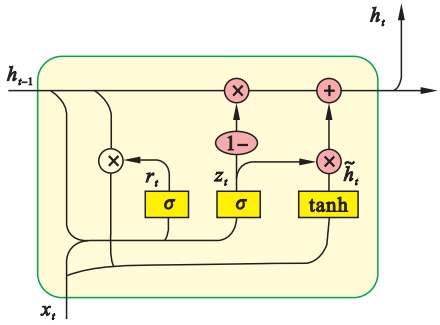

图 3 GRU 网络结构模型

Fig. 3 GRU network structure model

2.2 模型框架

本文提出的 LS – GRU 模型采用 LSTM + GRU + Self – attention 的架构. 如图 4 所示,LS – GRU 模型分为 4 层:第一层为拼接层,将所有词向量拼接为句子向量,句子向量拼接为整个影像描述段落向量;第二层为特征预提取,经过一层 LSTM 网络,将段落进行预筛查获取;第三层为 GRU 深度分析层,该层主要负责对文本特征的学习与提取,将 LSTM 网络预处理得到的特征进行深度学习;第四层为自注意力层,用来定位关键特征信息,如血管床模糊、高密度影等重要特征信息.

本文选择将句子作为神经网络训练的基本单元,由拼接层输入到 LSTM 层的数据为

$$\boldsymbol{X} = \boldsymbol{N}_{\mathrm{BS}} \times \boldsymbol{N}_{\mathrm{PL}} \times \boldsymbol{N}_{\mathrm{SL}} \times L_{\mathrm{WV}}. \tag{1}$$

其中:$\boldsymbol{X}$ 为输入网络的文档数据;$\boldsymbol{N}_{\mathrm{BS}}$ 为一次输入训练的文档个数;$\boldsymbol{N}_{\mathrm{PL}}$ 为段落长度,即句子的个数;$\boldsymbol{N}_{\mathrm{SL}}$ 为句子长度,即词的个数;L_{WV} 为词的长度,即词向量的长度. 最终输入预提取层数据为一个 $[\boldsymbol{N}_{\mathrm{BS}};\boldsymbol{N}_{\mathrm{PL}};\ \boldsymbol{N}_{\mathrm{SL}};\ L_{\mathrm{WV}}]$ 的矩阵.

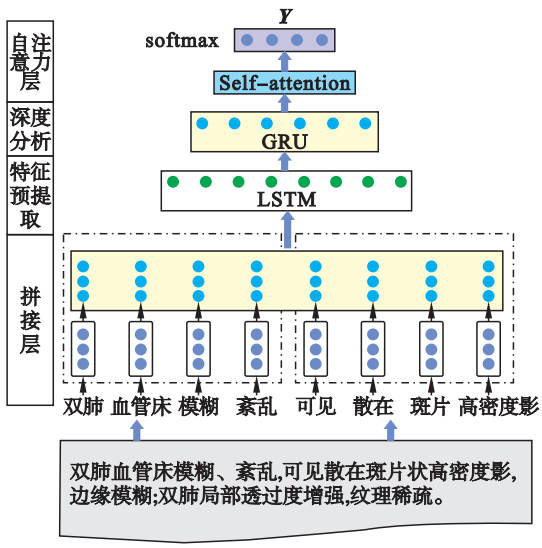

图 4 LS – GRU 网络结构模型

Fig. 4 LS-GRU network structure model

输出层 $\boldsymbol{Y}$ 是一个 $[\boldsymbol{N}_{\mathrm{BS}},2]$ 大小的矩阵,文中采用 onehot 编码方式,分类结果为 [0,1] 和 [1,0],分别代表了慢阻肺和肺炎两种临床慢性病. 对于预测结果和真实结果之间的判断采用交叉熵计算. 深度学习优化器选择 Adam 优化算法,并且引入了二次方梯度矫正方法,能够在一定程度上避免出现震荡和梯度消失.

自注意力层:查询矩阵、键矩阵和值矩阵分别用 $\boldsymbol{Q}$,$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 表示,这 3 个矩阵均来自最后一层 GRU 网络的输出 $\boldsymbol{h}_0$,整个自注意力计算过程如图 5 所示:首先计算 $\boldsymbol{Q}$ 与 $\boldsymbol{K}$ 之间的点乘,在计算过程中为了防止其结果过大,会用一个尺度标度 $\sqrt{d_k}$ 进行规范,其中 $\sqrt{d_k}$ 为 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 矩阵的维度. 最终结果归一化概率分布处理,采用 softmax 操作,然后再与矩阵 $\boldsymbol{V}$ 相乘得到最终结果表示. 该

操作可以表示为

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}. \quad (2)$$

$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$ 用来做最终文本分类. 该方法可以计算句子中其中一词与其他所有词的相关度,从而将目标定位到感兴趣的词.

文中分类算法的性能通常采用准确率进行测评,计算公式如下:

$$P = \frac{\text{正确分类出的文本数量}}{\text{所有参与训练的文本数量}} \times 100\% . \quad (3)$$

准确率越高,算法分类效果越好.

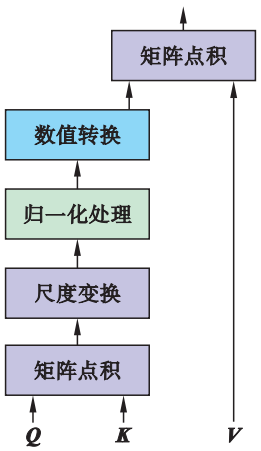

图 5　自注意力层结构示意图

Fig. 5　Self-attention structure diagram

3　实　　验

3.1　模型训练

在模型训练时,本文分别基于词向量分类模型和基于句子向量分类模型进行了训练和测试. 基于词向量的分类模型是指训练时将文档分割成一系列词,并将词输入深度神经网络训练的方法;基于句子向量分类模型是指将词向量以句子形式作为整体输入深度神经网络中进行训练的方法.

在计算词向量时对特殊字符进行筛选,删除了重复标记、乱码等特殊字符,然后使用 jieba 分词工具进行分词. 句子向量是将词向量进行拼接成固定长度,送入网络进行训练.

本文代码基于 tensorflow1.9,并且在具有 NVIDIA GeForce GTX 2070 的服务器上进行训练,单模型训练时占用内存约 6.7 GB,约需训练 19 h. 模型训练过程中使用的超参数如表 2 所示.

模型参数设置:输入为 50 个段落影像描述,经过拼接层,对段落进行分词、分句,最后进行规整化处理得到[50,48,27,30]矩阵训练作为提取层输入,50 为 Bath_size 大小,48 为段落中句子最大长度,27 为句子中词的最大个数,30 为 Word_embedding_dim 词向量维度;0.001 为 Learning_rate 模型学习率,0.5 为 Dropout 深度网络神经元丢弃的概率,即随机失活率. 预提取层采用 LSTM 结构,隐含单元设置为 50,输出为[50,48,27,50],同时将其作为 GRU 输入. 经过第三层的深度分析层后,LS－GRU 网络输出变为[50,48,50],并将该结果作为图 5 的查询、键和值输入到自注意力层(Self－attention)进行关键信息定位处理. 自注意力层输出向量为[50,48,50],输入 softmax 可得到[50,2]结果矩阵.

表 2　超参数设置

Table 2　Super parameter setting

超参数	值
Word_embedding_dim	30
Learning_rate	0.001
Dropout	0.5
Bath_size	50
Hidden_size	10
Adam_gamma	0.2

注:Word_embedding_dim 为词向量维度;Learning_rate 为训练模型学习率;Dropout 为随机失活率;Bath_size 为一次送入网络训练的文本数量;Hidden_size 为模型 GRU 隐含层大小;Adam_gamma 为训练模型梯度下降率.

本文模型训练流程见表 3.

表 3　算法流程

Table 3　Flow chart of the algorithm

算法 1　LS－GRU 网络训练流程

输入:　$\boldsymbol{X}$ 为输入网络的文本数据;
　　$\boldsymbol{Y}$ 为输出层矩阵;
　　P 为初始模型准确率;
　　Epoch 为训练模型轮数.

输出:　训练后的网络模型.

```
iters = getIters(X);获取一轮训练的次数
for i = 1:Epoch
  for j = 1:iters
    x = getBathsize(X);获取一次训练的数据
    y,p* = model(x);计算文本预测标签和精度
    if p* > P then
      save model
    endif
  end for
end for
return model
```

3.2　实验结果及分析

本文实验记录了每一个模型在验证集和测试集的精度,如表 4 所示.

从表 4 中可以看出,①基于句子为单位训练的模型相比以词为单位训练的模型效果要好. 因为医疗影像报告的叙述简单、明确,不像新闻媒

体、对话等自然语言含有很多情感词. 同时医疗文本词与词之间也同样具有较强关联性,因此将句子作为整体相比单个词训练的网络较为准确. ②模型网络相比单一网络精确度要高. 其原因是由于在网络前端和后端分别加入 LSTM 和 Self – attention 后能够更加准确地定位关键信息.

表 4 模型对比

Table 4 Models comparison

模型名称	验证集精度	测试集精度
SVM	053	0. 51
LSTM	0. 52	0. 5
GRU	0. 54	0. 52
LSTM + GRU	0. 55	0. 56
Word + MultiBasicGRU	0. 60	0. 62
Sentence + BiGRU	0. 64	0. 65
Sentence + MultiBasicGRU	0. 63	0. 625
Sentence + MultiBasicLSTM	0. 645	0. 675
LS – GRU	0. 658	0. 733 3

注:SVM,LSTM,GRU 是单独以此为模型进行分类的网络;LSTM + GRU 模型是将两者结合的文本分类网络模型;Word + MultiBasicGRU 为以词向量为单位训练的基础多层 GRU 深度网络;Sentence + BiGRU 为以句子向量为训练单位的双向 GRU 网络;Sentence + MultiBasicGRU 为以句子为单位训练的多层 GRU 网络;Sentence + MultiBasicLSTM 为以句子为单位训练的多层 LSTM 网络;LS – GRU 为本文网络结构.

由图 6 可以看出,本文提出的方法相比其他模型收敛速度更快,且在更早时间达到最佳训练结果.

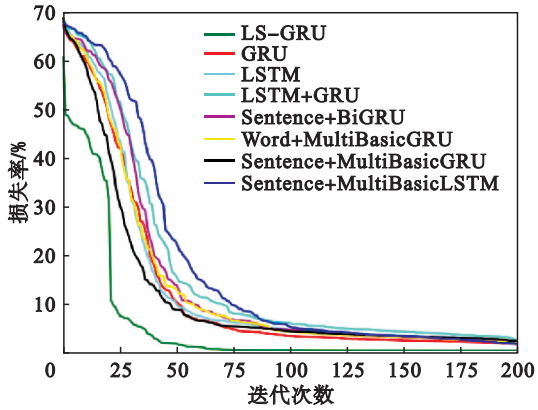

图 6 各训练模型损失函数曲线

Fig. 6 Loss function curves of training models

4 结 语

本文以 GRU 网络和 LSTM 网络结构为基础进行改进,提出了 LS – GRU 网络模型. 该模型对 1 168 例医疗影像报告进行文本分类获得了 0. 733 3 的分类精度. 本文提出的网络模型是在 GRU 网络前端和后端分别加入了 LSTM 和 Self_attention 结构,发现其精度相比单一网络要好,因此对数据的预处理和注意力机制有利于文本的分类.

同时也对其他网络结构进行了实验,基于双向 GRU 神经网络和双向 LSTM 网络的文本分类,虽然取得了类似于单一网络相当的结果,但网络复杂度相对较高,与本文网络结构相比占用的资源和时间都较大. 近年来胶囊网络在自然语言处理中得到很大发展,这将作为下一步研究的内容.

参考文献:

[1] Yu S, Indurthi S R, Back S, et al. A multi-stage memory augmented neural network for machine reading comprehension [C]//Proceedings of the Workshop on Machine Reading for Question Answering. Melbourne, 2018: 21 – 30.

[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems. Long Beach, 2017: 5998 – 6008.

[3] Lin Y, Shen S, Liu Z, et al. Neural relation extraction with selective attention over instances [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, 2016: 2124 – 2133.

[4] 陈志豪,刘子辰,邱大伟,等. 基于注意力和字嵌入的中文医疗问答匹配方法[J]. 计算机应用,2019,39(6):1639 – 1645.
(Chen Zhi-hao, Liu Zi-chen, Qiu Da-wei, et al. Chinese medical question-and-answer matching method based on attention and word embedding [J]. *Computer Application*, 2019, 39(6): 1639 – 1645.)

[5] 张浩宇,张鹏飞,李真真,等. 基于自注意力机制的阅读理解模型[J]. 中文信息学报,2018,32(12):125 – 131.
(Zhang Hao-yu, Zhang Peng-fei, Li Zhen-zhen, et al. Self-attention based machine reading comprehension [J]. *Journal of Chinese Information Processing*, 2018, 32(12): 125 – 131.)

[6] Kowsari K, Brown D E, Heidarysafa M, et al. Hdltex: hierarchical deep learning for text classification [C]//2017 16th IEEE International Conference on Machine Learning and Applications. Cancun: IEEE, 2017: 364 – 371.

[7] Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions [J]. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 1998, 6(2): 107 – 116.

[8] Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, 2016: 207 – 212.

[9] Huang E H, Socher R, Manning C D, et al. Improving word representations via global context and multiple word prototypes [C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju Island, 2012: 873 – 882.

(下转第 961 页)