基于深层结构模型的新词发现与情感倾向判定

孙 晓 孙重远 任福继
(合肥工业大学计算机与信息学院 合肥 230009)

摘 要 随着社交网络的发展,新的词汇不断出现。新词的出现往往表征了一定的社会热点,同时也代表了一定的公众情绪,新词的识别与情感倾向判定为公众情绪预测提供了一种新的思路。通过构建深层条件随机场模型进行序列标记,引入词性、单字位置和构词能力等特征,结合众包网络词典等第三方词典。传统的基于情感词典的方法难以对新词情感进行判定,基于神经网络的语言模型将单词表示为一个 K 维的词义向量,通过寻找新词词义向量空间中距离该新词最近的词,根据这些词的情感倾向以及与新词的词义距离,判断新词的情感倾向。通过在北京大学语料上的新词发现和情感倾向判定实验,验证了所提模型及方法的有效性,其中新词判断的 F 值为 0.991,情感识别准确率为 70%。

关键词 新词发现,条件随机场,深层结构模型,情感倾向判定,神经网络语言模型
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.9.040

New Word Detection and Emotional Tendency Judgment Based on Deep Structured Model

SUN Xiao SUN Chong-yuan REN Fu-ji
(School of Computer and Information, Hefei University of Technology, Hefei 230009, China)

Abstract With the development of social network, new words appear ceaselessly. The appearance of new word tends to characterize the social hot spot or represent certain public mood. The new word detection and emotional tendency judgment provide a new way for the public mood forecast. We constructed the deep conditional random fields model for the sequence labeling, introduced part of speech, character position, the ability of word formation as features, and combined it with the crowd sourcing network dictionary and the other third party dictionary. Traditional method based on emotional dictionary is difficult to judge the new word emotional tendency. We expressed word as a vector of K dimension based on neural network language model in order to find the nearest words to the new word in the vector space. According to the emotional tendency of these words and the distance between them and the new word, the new word sentiment is judged. The experiment on corpus of Peking university demonstrates the feasibility of the proposed model and method, in which the new word detection F-value is 0.991, and the emotion recognition accuracy is 70%.

Keywords New word detection, Conditional random fields, Deep structured model, Emotional tendency judgment, Neural network language model

1 引言

1.1 课题背景

随着社会进步和信息时代的到来,社交网络不断普及,新词的产生非常迅速。在网络中,用户不知道对方的各方面情况,自己也不为别人所了解,因此人们的言语往往会更加自由、随意,不会像书面语那样正规,例如"给力"、"坑爹"。由于微博等主流社会媒体的发表和评论的字数都是受到限制的,因此用户往往更乐于使用更短的词去表达更多的意思,例如"累觉不爱"、"人艰不拆"。大多数研究者认为新词与未登录词是一样的,这些词语都是没有在词典里出现的。每天在网络上都会出现一些代表着新观念的新词语,这些词语有的在一段时间内被广泛使用,而过了一段时间则迅速消失;有的则逐渐为人们所接受,成为了人们生活的一部分[1]。由于汉语没有限制词的空间并且中国的网络用户使用新词非常频繁,因此不可能有完备的词典包含所有的新词,这也就给中文分词带来了巨大的挑战。大量新词的产生严重影响到了分词的质量,有研究表明:60%的分词错误是由新词产生的[2]。而分词又是现实中很多应用领域的基础,所以新词的识别是自然语言发展的关键步骤。

互联网信息量是十分巨大的,并且每日仍在剧增,所以,要想在短时间内对人物、信息等提取有价值的信息是很困难的。新词的出现往往伴随着一定的社会热点或用户心情,新词是最近才出现的词,具有一定的代表性,其能够得到广泛传

到稿日期:2014-05-19 返修日期:2014-08-01 本文受国家自然科学基金项目(61203315),国家863计划(2012AA011103)资助。
孙 晓(1980—),男,副教授,硕士生导师,CCF会员,主要研究方向为自然语言处理、情感计算、机器学习算法等,E-mail:sunx@hfut.edu.cn;孙重远(1992—),男,硕士生,主要研究方向为自然语言处理、情感计算等;任福继(1959—),男,教授,博士生导师,主要研究方向为人工智能、情感计算。

播，要么是因为群众对这个词语比较感兴趣，乐于使用，要么就是这个词语能够最简单地表达群众的心情。这也为预测大众的情绪提供了一个新思路，即可以通过对最近的热点情感倾向判定来预测公众的情绪或对于某些事件的情感倾向。

1.2 相关工作

近年来新词的研究一直在不断发展，中科院计算所、哈尔滨工业大学、微软亚洲研究院等机构的研究人员都在新词识别领域开展了很多工作，并且取得了不错的成果[3]。

Fu[4]将未登录词的识别看作是一个分类问题，词的词性标注定义为它的类别，使用词性的上下文特征、词之间的连接模型和构词模式来对条件概率模型进行训练，寻找新词边界；同时还提出了对未登录词标注的解决方案，实验取得了不错的效果，但是预处理复杂，对语料的限制比较大。

Goh[5]使用 HMM 对词进行标记，用输出的标注结果训练 SVM 模型得到基于字符的标注，使用词块的字符序列对未登录词进行检测。结合未登录词的检测和初始分词结果，得到最终的分词结果。实验结果表明，该方法的检测结果还是比较令人满意的，但是在 HMM 粗分的时候容易引入误差。

Xu[6]先对语料进行分词和词性标注，然后从训练语料中提取出积极和消极的样本，对训练语料里各种类型的词语分类，通过 SVM 训练得到新词的空间向量。在测试语料上结合相关的约束和松弛变量预测候选新词，将候选新词向量化后输入 SVM 分类器，结合词本身特性来计算候新词的评价值，与相关的阈值相比较得到新词判定结果。

Li[7]提出了一种新词模式特征，从一些积极的和消极的训练语料中训练出新词的内部构词模式，用训练语料提取的词模式来量化新的 SVM 分类器，在包含新词的测试语料上进行 SVM 测试，最后通过规则过滤器得到最后的新词识别结果。该方法可以提高单词的识别率和召回率。

Zeng[8]为了弥补传统方法中的离线训练的缺陷，提出了一种改进的基于局部上下文信息预测部分匹配（PPM）的分词算法，这种算法侧重于在线分词和新词检测，在开放测试和封闭测试上都取得了很好的效果。但该方法复杂度较高，且执行时间较长。

陈飞[9]利用条件随机场（CRFs）将新词发现问题转化为预测已分词词语边界是否为新词边界的问题，提出了很多区分新词边界的统计特征，并采用 CRFs 方法综合这些特征，实现了开放领域新词发现的算法。

近年来对于新词的研究主要集中在新词的识别和词性判断上，而对于新词的情感判定及其应用的研究很少，新词的情感倾向对文本的情感倾向具体有多大的影响也无法得知。但是新词的情感判定为判断群众对新闻的情感以及群众的心理状态提供了一个新的方法和思路。目前对于新词的情感判定尚无针对性的解决办法。

张[10]使用机器学习对词语倾向性自动判定，特征向量采用互信息，通过训练朴素贝叶斯、支持向量机、基本规则、K 最近邻、决策树 5 种分类器并进行性能调优来进行自动属性判定。在 COAE 语料上的实验验证了使用机器学习进行主观词情感倾向判定的方法的有效性，但该方法的效果仍然受到种子词典质量的限制。

郑[11]使用构建独立词的“词袋模型”，并用深度神经网络将词转化为词向量，最后用 k-means 进行词义聚类；在构建模型之前，使用词库对原始文本进行筛选，并指定了 4 条上下文规则来选择出那些没有被词库识别出来的词，并将相同的词聚集到一起，取得了不错的效果。

1.3 本文主要研究内容

新词识别是中文自然语言领域的基础，而新词的情感也为预测大众情感提供了一个新的思路。本文的主要研究内容如下：

（1）构建条件随机场模型（CRFs）进行新词发现，在线性条件随机场的基础上，构建了两层条件随机场，并引入了众包网络词典作为第三方词典，单字词性标记、单字构词能力等作为发现新词的特征。

（2）构建神经网络语言模型进行新词情感判定工作，通过迭代包含新词的语料，将语料中的词语转化为词义向量，通过寻找词义向量的相似度来比较文本的相似度，寻找与新词最接近的一些词，再进行基于词典的情感判定，使用的词典为台湾大学的正负面情感词典，判定出这些近义词的情感，从而判定出新词的情感倾向。

（3）提出了一种新词训练和识别框架，为了构建更自然的新词训练语料，利用以时间序列产生的社会新闻语料（前几个月做基础词典，后几个月做新词训练和测试语料），可以构建更准确的新词识别框架，语料中未在系统词典中出现的词均标记为新词，这样做使得新词的范围更广泛，且标记出的新词特征和分布与真实社交网络文本中的新词更加相符。

（4）为提高新词识别的准确率，引入众包网络词典中的单字在词中的位置信息作为特征，Webdict 网络词典是一个以众包方式进行不断更新的网络词典。本文统计每一个字出现的总次数和在词头、词中、词尾的出现次数，将其作为新词识别特征，有助于判断多个字的组合能否构成新词。

2 模型方法介绍

2.1 深层 CRFs 模型

2.1.1 线性链 CRFs 模型

条件随机场（Conditional Random Fields，CRFs）是在整个观察序列上直接估计状态序列概率的判别模型，相比生成模型，例如隐马尔科夫（Hidden Markov Model，HMM）模型，避免了标记偏见的问题。CRFs 已被广泛应用于解决序列标记问题。序列标记问题比较常用的线性链 CRFs 结构如图 1 所示，是 CRFs 的一个高效实现，且结构简单。

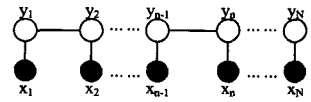

图 1 线性链 CRFs 的图形表示

$x(x_1, x_2, x_3, \cdots, x_N)$ 表示所给定的一个观察序列，状态序列为 $y(y_1, y_2, y_3, \cdots, y_N)$。

$t_j(y_{n-1}, y_n, x, n)$：观察序列标记位置 $n-1$ 与 n 之间的转移特征函数。

$s_k(y_n, x, n)$：观察序列 n 位置的状态特征函数。

将两个函数统一为：$f_m(y_{n-1}, y_n, x, n)$，根据随机场基本理论：

$$p(y|x,\Delta)=\frac{\exp(\sum_{n,m}\lambda_m f_m(y_{n-1},y_n,x,n))}{M(x;\Delta)} \quad (1)$$

$$M(x;\Delta)=\sum_{y}\exp(\sum_{n,m}\lambda_m f_m(y_{n-1},y_n,x,n)) \quad (2)$$

模型参数 $\Delta=\{\lambda_i\}$ 是 $L2$ 正则条件下状态序列最大似然参数集。得到模型参数后，可以利用 CRFs 来对输入的句子序列进行新词位置角色标注，从而得到句子中的新词。

2.1.2 深层 CRFs 模型

通过将线性链 CRFs 模型堆叠在一起，构建深层结构的 CRFs 模型，可以达到更高的分类精度，更适合在新词发现中发现新词的隐藏结构特征。多层结构的 CRFs 模型是 Dong Yu 和 Li Deng 针对语音信号处理而提出的[12]，本文将其改进并引入到文本处理中。本文采用两层的 CRFs 模型，结构如图 2 所示。

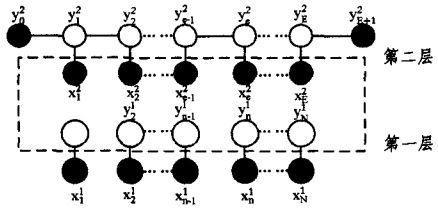

图 2 深层 CRFs 的图形表示

第一层是一个不使用状态转换特征的零阶 CRFs，第二层是一个线性链 CRFs(加入了开始状态和结束状态)。在第一层使用零阶 CRFs 减少了计算量，提高了效率。在深层 CRFs 中，第 i 层的观察序列由两部分组成：前一层的观察序列 x^{i-1} 和分量级别的边界后验概率 $P(y_n^{i-1}|x^{i-1})$。在观察值上的特征构造只使用输入信息的一部分。在深层 CRFs 模型中状态序列预测是自底向上逐层进行的，因此计算复杂度被限制在使用层数的线性范围内。深层 CRFs 的参数估计比线性 CRFs 更为复杂，采用逐层监督学习的训练方法，对两层 CRFs 模型进行参数估计[13]。两层 CRFs 模型输出的是一个状态序列，第二层的参数是通过状态序列水平的正则条件对数似然最大化来进行优化的，第一层是使用如下边缘概率对数似然的最大化来进行训练的：

$$J(\Delta,X)=\sum_{k,n}\log p(y_n^{(K)}|x^{(k)};\Delta)-\frac{\|\Delta\|^2}{2\sigma^2} \quad (3)$$

此边缘概率是从第一层传向第二层的唯一信息，第二层 CRFs 基于该边缘概率和观察序列进行训练。$J(\Delta,X)$ 可以以 $O(TY)$ 复杂度优化，Y 是状态数，T 是维数。

2.2 神经网络语言模型

神经网络语言模型(Neural Network Language Model，NNLM)的结构如图 3 所示。这是一个标准的 3 层前向神经网络，最下方的 $W_{t-n+1}\cdots W_{t-1}$ 是要预测下一个词 W_t 的前 $n-1$ 个词。C 为将词映射到初始词向量的映射，$C(n)$ 即为词 n 的词向量。输入层为这些单词的词向量，将这些词向量首尾相连成为一个新的向量 x 作为第二层隐藏层的输入。隐藏层为普通的神经网络，设 D 为一个偏置项，直接用 $D+Hx$ 得到输出，激活函数为 $\tanh$。第三层为输出层，有 $|V|$ 个节点，V 为语料中的词数，节点 y_i 表示下一个为 i 的 log 概率，最后对其做 soft max，将输出值 y 做归一化。

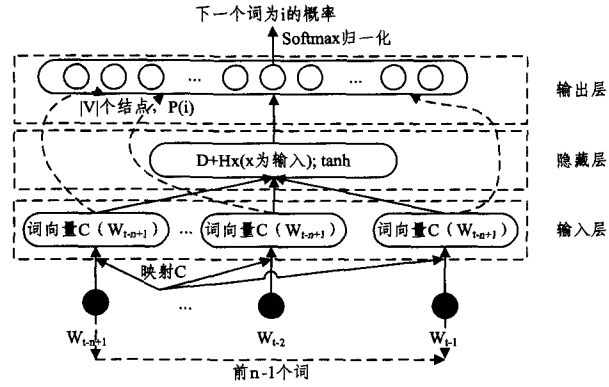

图 3 神经网络语言模型

通过寻找使得训练语料惩罚对数似然值最大化的 (θ,C) 来完成语料的训练，使用随机梯度下降法来进行模型的优化，得到语言模型和词向量。

若新词的词义向量为 $A(a_1,a_2,a_3,\cdots,a_n)$，某词的词义向量为 $B(b_1,b_2,b_3,\cdots,b_n)$，新词与该词的相似度定义为 A 与 B 夹角的余弦值，为

$$\cos\langle A,B\rangle=\frac{A\times B}{|A|\times|B|} \quad (4)$$

计算每个词与新词的距离，寻找出与新词最接近的一些词，通过这些词的情感倾向来判定新词的情感倾向。

3 实验与结果

3.1 新词发现实验

3.1.1 特征选择与语料处理

条件随机场是在训练集和测试集的多个特征上进行学习和预测的，本文引入基于单字的特征，根据这些单字的特征来挖掘该字构造新词的可能性。利用提出的新词训练和识别框架得到训练语料和测试语料，具体为：对以时间序列产生的北大人民日报前 6 个月的新闻语料进行预处理、词频统计和筛选后构建系统词典，作为判定新词的依据。使用 7 月份的语料作为训练语料，8 月份的语料作为测试语料。进行单字的词性位置标记，虽然词一级的词性标注能给出一些词间信息，但是给出的信息是有限的，本文对单字进行词性位置标记，以挖掘更深层次的构词特征。语料中的单字词性标记例子如表 1 所列。按照 4-tag 的标记方式，对词性进行了二次位置标记。

表 1 单字词性位置标记

李	nr-o	约	ns-e	马	n-b
长	nr-b	集	n-b	车	n-e
圣	nr-e	团	n-e	”	w-o
会	v-b	“	w-o	外	n-b
见	v-e	三	m-o	长	n-e
李	ns-b	驾	q-o	。	w-o

引入 Webdict 网络众包词典作为第三方词典，计算单字的构词能力，标注出单字在 Webdict 词表中出现的总次数以及分别在 B、I、E 位置上出现的次数。对训练与测试语料中各个字计算特征取值，再对词典进行匹配。本文在语料准备中对传统的四标记集 (B,I,E,O) 进行了改进，更有利于新词发现实验。B 表示这个汉字作为新词的第一个字出现，I 表示这个汉字作为三字及以上新词的中间字出现，E 表示这个汉字作为新词的最后一个字出现，O 表示这个汉字在已登录词中出现，本文将在系统词典中出现的词标记为已登录词，将

未在系统词典中出现的词标记为新词。标注后的实例如表 2 所列。第 1 列为语料中的字，第 2 列为该字的特征取值，包括单字词性位置特征和单字构词能力特征取值，第 3 列为该字的正确标记。

表 2　训练语料标注后实例

字序列	特征值	标记
克	nz-b 957 163 531 263	B
隆	nz-e 185 54 67 64	E
技	n-o 558 90 412 56	O
术	n-o 559 9 422 168	O
的	u-o 296 19 119 158	O
研	n-b 172 73 71 28	B
究	n-e 84 7 54 23	E

3.1.2　特征模版

使用 CRFs 进行新词发现的特征模板如表 3 所列。C_0 表示当前字，负数表示前面的字，正数表示后面的字，连在一起表示多个字的组合特征。

表 3　特征模版

类型	特征模版	描述
一元	C_{-2}，C_{-1}，C_0，C_1，C_2	单字特征
二元	$C_{-2}C_{-1}$，$C_{-1}C_0$，C_0C_1，C_1C_2	双字组合特征
三元	$C_{-2}C_{-1}C_0$，$C_{-1}C_0C_1$，$C_0C_1C_2$	三字组合特征
系统词典	$S(C_0)$，$S(C_{-1}C_0)$，$S(C_0C_1)$	是否存在于系统词典词中
第三方词典	$T(C_0)$，$T(C_{-1}C_0)$，$T(C_0C_1)$	是否存在于第三方词典词中

3.1.3　衡量标准

本文采用正确率(P)、召回率(R)和 F 值来对实验结果进行衡量，具体定义为：

$$P=\frac{\text{正确识别的新词个数}}{\text{识别的新词个数}} \tag{5}$$

$$R=\frac{\text{正确识别的新词个数}}{\text{应被识别的新词个数}} \tag{6}$$

$$F=\frac{2\times P\times R}{P+R} \tag{7}$$

3.1.4　实验结果及分析

使用 conlleval. pl 程序对实验结果进行测评，表 4 所列为使用单层 CRFs 分别在词性标注以及本文引入的单字词性位置标记和构词能力的特征下的实验结果。

表 4　单层 CRFs 实验结果

特征选择	正确率(P)	召回率(R)	F 值
词性标注	0.939	0.935	0.937
单字词性位置＋众包词典构词能力	0.993	0.986	0.989

从实验结果中可以看出，本实验中的准确率和召回率都比较高，出现这样的结果主要有以下方面原因：本文使用了前 6 个月人民日报语料分词后构建的词典，而未在词典中出现的被定义为新词，扩大了新词的范围，使得语料的可靠性更高；实验的语料为人民日报的新闻文章，无论是在汉语用词还是语法结构上都要比互联网文档和微博语料更正式、规范，领域也比后者窄，因此在识别的过程中噪声明显会比其他语料小；训练语料和测试语料分别为一个月的人民日报语料，语料包含约 225 万字，36 万词，在这种比较大规模的语料下，上下文和词性标注提供了比较多的信息，所以识别的效果会更好。

在引入了词性位置标记和单字构词能力后，准确率和召回率得到了大幅的升高，分别提升了 6 个和 5 个百分点，主要由于在词性标记位置和 Webdict 网络词表中的位置信息包含了大量信息，前者让 CRFs 学习到在某个字通常以某个词性出现的时候，是在什么位置；而后者则让 CRFs 学习到字在各个位置出现的信息熵，例如在 B 位置出现的很多词的字在构建新词的时候，往往更容易出现在新词的开头。

对两层 CRFs 使用相同的语料进行实验，结果如表 5 所列。

表 5　深层 CRFs 实验结果

特征选择	正确率(P)	召回率(R)	F 值
词性标注	0.952	0.944	0.948
单字词性位置＋众包词典构词能力	0.994	0.988	0.991

在引入了两层 CRFs 之后，准确率和召回率也得到了一定的提升，这是由于第二层 CRFs 的特征选择是在第一层的基础上进行的。语料的规模本身也比较大，引入深层后更能够刻画特征的内在信息，通过样本特征空间的变换，从而使预测更加准确。从数据可以看出：在引入了本实验的特征后，深层 CRFs 比单层 CRFs 的 F 值仅上升了 0.2 个百分点，这是因为使用的人民日报语料的噪音较小，根据单字词性位置和构词特征已经能够学习到比较好的模型，因此引入深层 CRFs 后提升效果不是很明显，预测在微博和互联网等语料上效果会更好，这有待进一步的实验。

图 4 所示为深层 CRFs 与单层 CRFs 的结果对比。在图中，分别对 4 种实验下的准确率、召回率和 F 值进行了比较，用不同的灰度表示不同的模型和特征。

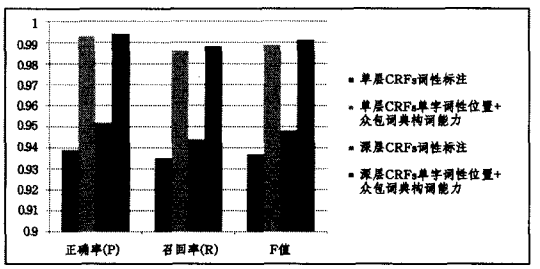

图 4　实验结果对比

选取本次实验中的一篇新闻报道为例，实验结果如图 5 所示。列出了发现的一些新词，其中新词用下划线表示。可以看出，即使出现次数很少的新词也可以被识别出来。

据<<日本经济新闻>>报道，这一科技信息处理系统将被命名为“ITBL”，意思是“以信息技术为基础的研究室”。它将把全国三所国立大学和科研机构的6台超级计算机，用传输速度为每秒24亿比特的高速通信线路连接起来，使不同科研单位的研究人员都能利用这个信息处理系统，从事基因、蛋白质、原子、分子等各种微观结构的研究工作。

图 5　新词识别例子

3.2　新词情感倾向判定实验

3.2.1　实验结果

本文使用 8 月份人民日报语料作为训练语料，训练出词义向量，并通过向量间的距离来表征词语间的距离。选取与新词词义最接近的 10 个词，使用台湾大学正负面情感词典进行匹配，来识别新词的情感。随机选取 10 个新词进行情感判定测试，测试结果如表 6 所列，表中列出了新词、测试结果中

与新词最接近的5个词以及根据这些词对新词进行情感倾向判定的结果。

表6 新词情感倾向判定抽样实验结果

新词	近义词	情感倾向
善意	玩笑话、执掌、敬爱、世人、约翰·布拉格	正面情感
道听途说	胡编、耳光、何故、难于、超然物外	中性情感
听话	家伙、豆汁儿、心领神会、随口、大人	负面情感
吃喝嫖赌	菜叶、澳门、防护兵、夜生活、浴场	负面情感
巴西队	比分、胜局、朝鲜队、战平、小将	正面情感
正版品	库存、销售量、纯利、喜讯、平均价	中性情感
盗窃案	厂房、郊区、站台、辱骂、刑讯逼供	负面情感
老龄化	估量、全球化、产物、必然性、基本矛盾	负面情感
感染力	文学、应和、歌唱、魅力、评论家	正面情感
打击报复	控告人、检举、堵塞、泄漏、举报人	负面情感

3.2.2 实验结果分析

对结果进行了人为判定，随机抽取的10个新词中7个情感判定为正确的，准确率为70%。由于是随机抽取的，因此只是对本方法进行了一个简单的测试，从测试结果中可以看出本方法是有效可行的。如表6所列，根据新词词义向量选出最相似的5个词中多数是与新词相似或相关的，根据这些词来推测新词的情感是具有一定代表性的。

3.2.3 错误分析

分析判定错误的新词，总结新词产生的主要原因，人为判定实验包括3个错误，分别为：将“道听途说”判定为中性情感，应为负面情感；将“听话”判定为负面情感，应为正面情感；将“巴西队”判定为正面情感，应为中性情感。错误的原因主要是汉语词语在某些特定的情形下，会有特定的情感。例如本实验中的“巴西队”，通常情况下，巴西队代表的是一个国家的足球队，是没有情感的，但是在社交网络中，大多数时候“巴西队”的出现是伴随着“巴西队”比赛的胜利或者“巴西队”球星的活动，因此人们对于这些常常是有正面情感的；再例如本实验中的“听话”，一般是形容一个人听从长辈或领导的话，顺从长辈或领导的意志，应该为一个正面情感的词语，但是少数情况下，又会用来讽刺一个人没有自己的主见。因此要解决这类问题，最好的办法就是使用更大规模的语料，语料中包含较多的新词，迭代出的词义向量就能更好地反映该词的通俗词义。按照本方法可以进行更准确的新词情感倾向判定。

结束语 随着社交网络的迅速普及，社交网络上的信息呈现出指数级增长的趋势。在社交网络中，因为其文本具有口语化的特点，产生了很多新词和新表达方式。新词往往是伴随着一些公共事件产生，而且部分包含了公众一定的情感倾向。新词是不存在于系统词典中的词汇，其产生虽然有一定的构词规律，但自由度较大，因此，对新词进行准确识别，进而对其情感倾向进行判断，是社交网络时代中文信息处理的关键和难点。本文提出了一种新词训练和识别框架，为了构建更自然的新词训练语料，利用以时间序列产生的社会新闻语料（前几个月的作为基础词典，后几个月的作为新词训练和测试语料），可以构建更准确的新词识别框架；另外，在该语料上通过神经网络语言模型对识别后的新词进行迭代，将新词用上下文向量来表示其词义，进而基于词义向量空间距离对新词的情感进行判断。论文还引入了以众包方式构建的网络新词词典，通过引入单字在该词典中的位置特征，来进一步提高系统的鲁棒性。论文将用于语音信号处理的多层CRFs改进并用于新词序列标记，改进后的多层模型相对线性CRFs具有更好的学习隐藏结构的能力，更适合新词识别任务。通过在北大语料库上的实验证明了本方法和模型的有效性，进一步将在微博和互联网网页语料上开展相关工作。

参 考 文 献

[1] 聂金慧，苏红旗，时志远. 中文新词提取与过滤研究综述[J]. 中国科技博览，2013(30)：209-210
Nie Jin-hui, Su Hong-qi, Shi Zhi-yuan. Survey of Chinese new words extracting and filtering[J]. China Science and Technology Review, 2013(30): 209-210

[2] Sproat R, Emerson T. The First International Chinese Word Segmentation Bakeoff[C]// Proceedings of the Second SIGHAN Workshop on Chinese Language Processing. Sapporo, Japan, 2003: 133-143

[3] 张海军，史树敏，朱朝勇，等. 中文新词识别技术综述[J]. 计算机科学，2010，37(3)：6-10
Zhang Hai-jun, Shi Shu-min, Zhu Chao-yong, et al. Survey of Chinese new words identification[J]. Computer science, 2010, 37(3): 6-10

[4] Fu G, Luke K-k. Chinese Unknown Word Identification Using-Class based LM [C]// Proceedings of The First International Joint Conference on Natural Language Processing. Hainan Island, China, 2004: 262-269

[5] Goh C-L, Asahara M, Matsumoto Y. Machine Learning-based Methods to Chinese Unknown Word Detection and POS Tag Guessing[J]. Journal of Chinese Language and Computing, 2006, 16(4): 185-206

[6] Xu Yuan-fang, Gu Hui. New Word Recognition Based On Support Vector Machines And Constraints[C]// Proceedings of 2013 IEEE International Conference on Computer Science and Automation Engineering. Singapore, 2013: 56-59

[7] Li Cheng-cheng, Xu Yuan-fang. Using on support vector and wordfeatures new word discovery research[M]// Trustworthy Computing and Services. Springer Berlin Heidelberg, 2013: 287-294

[8] Zeng Hua-lin, Zhou Chang-le, Zheng Xu-ling. A New Word Detection Method for Chinese based on local context information [J]. Journal of Donghua University(English version), 2010, 27(2): 189-192

[9] 陈飞，刘奕群，魏超，等. 基于条件随机场方法的开放领域新词发现[J]. 软件学报，2013，24(5)：1051-1060
Chen Fei, Liu Yi-qun, Wei Chao, et al. Open Domain New Word-Detection Based on Condition Random Field Method[J]. Journal of Software, 2013, 24(5): 1051-1060

[10] 张靖，金浩. 汉语词语情感倾向自动判断研究[J]. 计算机工程，2010，36(23)：194-196
Zhang Jing, Jin Hao. Study on Chinese word sentiment Polarity Automatic. Estimation [J]. Computer Engineering, 2010, 36(23): 194-196

[11] 郑文超，徐鹏. 利用word2vec对中文词进行聚类的研究[J]. 软件，2013，34(12)：160-162
Zheng Wen-chao, Xu Peng. Research on Chinese words Clustering with word2vec[J]. Computer Engineering and Software, 2013, 34(12): 160-162

[12] Dong Yu, Li Deng, Wang Shi-zhen. Learning in the deep-structured conditional random fields[C]// Proc. NIPS Workshop. 2009:1-8

[13] Peng Fu-chun, Feng Fang-fang, McCallum A. Chinese segmentation and new word detection using conditional random fields[C]// Proceedings of the 20th International Conference on Computational Linguistics. 2004:562-568

[14] 邱泉清,苗夺谦,张志飞. 中文微博命名实体识别[J]. 计算机科学,2013,40(6):196-198
Qiu Quan-qing, Miao Duo-qian, Zhang Zhi-fei. Named entity recognition on Chinese micro-blog [J]. Computer science, 2013, 40(6):196-198

[15] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013

[16] Xu Wei, Rudnicky A. Can artificial neural networks learn language models? [C]// The Proceedings of the 6th International Conference on Spoken Language Processing. 2000:202-205

(上接第 203 页)

缩短了其居民到达城市中心 CBD 和其他商业、休闲中心区域的时间,因此这种交通便利性的提升也就反映到了房价的增长上。另外,作为长沙市轨道交通从无到有的标志,地铁 2 号线的开通同时也推动了尚在建设中的地铁 1 号线地铁站点周边的房价上涨,表明市场已经开始充分认识到轨道交通因素对住宅价格带来的增值效应。

从另一方面来看,城市中心区域的地铁站点周边房价却出现下跌。这一现象不难解释:对于处于传统城市中心区域的住宅来说,其交通便利性已经很高。地铁的开通对其交通便利性的提升较小,因此其对应的房价提升空间也非常有限。而相较于轨道交通因素促发的其他区域住宅价格的显著增值,城市中心的住宅价格却相对下降。正是由于轨道交通的出现,从城市外围区域进入城市中心区域的成本降低,居民的生活和出行区域开始从单一的城市中心 CBD 向外围地铁沿线延展,从而也相对弱化了城市中心区域的地铁站点对周边住宅价格的影响。

结束语 本文基于网络爬虫获取的真实 Web 数据,以长沙地铁 2 号线工程为例,分析了地铁因素对周边房价造成的时空的影响,得出以下结论:(1)住宅房产越靠近地铁站点价格越高,距离地铁站点的距离每增加 1000 米,价格下降 1.9%;(2)地铁站点影响周边住宅价格的最显著范围为 2000 米;(3)地铁开通后,城市中心区域的地铁站点周边住宅价格下降,城市外围区域的地铁站点周边住宅价格上升。以上结论可作为消费者和房地产机构进行购房、投资和规划的参考。

参 考 文 献

[1] Chen W Y, Jim C Y. Amenities and disamenities: a hedonic analysis of the heterogeneous urban landscape in Shenzhen (China) [J]. The Geographical Journal, 2010, 176(3):227-240

[2] Holly S, Pesaran MH, Yamagata T. A spatio-temporal model of house prices in the USA [J]. Journal of Econometrics, 2010, 158(1):160-173

[3] 冯长春,李维瑄,赵蕃蕃. 轨道交通对其沿线商品住宅价格的影响分析——以北京地铁 5 号线为例[J]. 地理学报,2011,66(8):1055-1062
Feng Chang-chun, Li Wei-xuan, Zhao Fan-fan. Influence of rail transit on nearby commodity housing prices: a case study of Beijing Subway Line Five [J]. Acta Geographica Sinica, 2011, 66(8):1055-1062

[4] Wu J, Deng Y. House price index construction in the nascent housing market: the case of China [J]. The Journal of Real Estate Finance and Economics, 2014, 48(3):522-545

[5] Holly S, Pesaran M H. The spatial and temporal diffusion of house prices in the UK [J]. Journal of Econometrics, 2011, 69(1):2-23

[6] Case B, Clapp J, Dubin R, et al. Modeling spatial and temporal house price patterns: a comparison of four models [J]. The Journal of Real Estate Finance and Economics, 2004, 29(2):167-191

[7] Zheng S, Kahn ME. Land and residential property markets in a booming economy: New evidence from Beijing[J]. Journal of Urban Economics, 2008, 63(2):743-757

[8] 杨鸿. 城市轨道交通对住宅价格影响的理论与实证研究——以杭州地铁为例[D]. 杭州:浙江大学,2010
Yang Hong. Effects of urban rail transit on housing prices: a case study of Hangzhou subway [D]. Hangzhou: Zhejiang University, 2010

[9] Huang Hao, Yin Li. Creating sustainable urban built environments: An application of hedonic house price models in Wuhan, China [J]. Journal of Housing and the Built Environment, 2014, 30(4):1566-4910

[10] Lin Jen-jia, Hwang Chi-hau. Analysis of property prices before and after the opening of the Taipei subway system [J]. The Annals of Regional Science, 2004, 38(4):687-704

[11] 梅志雄,徐颂军,欧阳军,等. 广州地铁三号线对周边住宅价格的时空影响效应[J]. 地理科学,2011,31(7):836-842
Mei Zhi-xiong, Xu Song-jun, Ouyang Jun, et al. Spatio-temporal impact effects of Guangzhou Metro 3rd Line on housing prices [J]. Scientia Geographica Sinica, 2011, 31(7):836-842

[12] 郑捷奋,刘洪玉. 深圳地铁建设对站点周边住宅价值的影响[J]. 铁道学报,2005,27(5):11-18
Zheng Jie-fen, Liu Hong-yu. The impact of URRT on house prices in Shenzhen [J]. Journal of the China Railway Society, 2005, 27(5):11-18

[13] 聂冲,温海珍,樊晓锋. 城市轨道交通对房地产增值的时空效应[J]. 地理研究,2010,29(5):801-810
Nie Chong, Wen Hai-zhen, Fan Xiao-feng. The spacial and temporal effect on property value increment with the development of urban rapid transit: an empirical research [J]. Geographical Research, 2010, 29(5):801-810

[14] Knaap G J, Ding C, et al. Do plans matter? The effects of light rail plans on land values in station areas [J]. Journal of Planning Education and Research, 2001, 21(1):32-39

[15] 李汪. 长沙市房地产特征价格研究[D]. 长沙:湖南大学,2008
Li Wang. Research on the hedonic prices method of real estate in Changsha city [D]. Changsha: Hunan University, 2008