第47卷第1期 中国科学技术大学学报 Vol. 47, No. 1
2017年1月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jan. 2017

文章编号:0253-2778(2017)1-0063-07

# 利用新词探测提高中文微博的情感表达抽取

万 琪[1],于中华[1],陈 黎[1],宋磊磊[1],丁革建[2]

(1. 四川大学计算机学院,四川成都 610065; 2. 浙江师范大学数理与信息工程学院,浙江金华 321004)

**摘要**:情感表达抽取工作是细粒度情感挖掘的重要任务之一.中文微博中包含大量网络新词和不规范词,现有的方法在进行微博情感表达抽取任务时不能很好地处理上述情况.通过研究发现,微博中新词大量分布在文本的情感表达部分,于是提出了基于CRF的联合抽取模型,即将新词发现融入到情感表达抽取任务中,从而改进原有工作的不足.实验结果表明,新词探测对微博文本情感表达抽取有很好的指示作用,在电影领域和开放领域的微博数据集上分别进行实验,$F_1$ 值均提高了2%以上.

**关键词**:情感分析;新词发现;条件随机场;信息抽取

**中图分类号**:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2017.01.009

**引用格式**:万琪,于中华,陈黎,等. 利用新词探测提高中文微博的情感表达抽取[J]. 中国科学技术大学学报,2017,47(1):63-69.

WAN Qi, YU Zhonghua, CHEN Li, et al. Improving emotion expression extraction in Chinese microblogs via new words detection[J]. Journal of University of Science and Technology of China, 2017,47(1):63-69.

# Improving emotion expression extraction in Chinese microblogs via new words detection

WAN Qi[1], YU Zhonghua[1], CHEN Li[1], SONG Leilei[1], DIN Gejian[2]

(1. *College of Computer Science, Sichuan University, Chengdu* 610065, *China*;
2. *College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, Jinghua* 321004, *China*)

**Abstract**: Emotion expression extraction is one of the important tasks of fine-grained sentiment mining. Existing methods lack efficiency in dealing with this task in Chinese microblogs because there are many new words and non-standard words in them. It's found in this paper that a large number of new words are distributed in emotional expressions of the text in Chinese microblogs. A combined extraction model based on CRF is proposed, which incorporates new word detection into the task to improve the original work. The experimental results show that new word detection has good correlation with emotion expression extraction from Chinese microblogs, and that $F_1$ value increases more than 2% on both the data sets of the movie field and the open field in Chinese microblogs.

**Key words**: sentiment analysis; new word detection; conditional random field; information extraction

**收稿日期**:2016-03-01;**修回日期**:2016-09-17

**基金项目**:四川省科技支撑项目(2014GZ0063),浙江省自然科学基金(LY12F02010)资助.

**作者简介**:万琪,男,1991年生,硕士生.研究方向:自然语言处理. E-mail:youngwq12@163.com

**通讯作者**:于中华,博士/副教授. E-mail: yuzhonghua@scu.edu.cn

## 0 引言

随着中文社交网络的兴起，微博文本的情感挖掘工作成为目前研究的热点.用户在社交媒体上对感兴趣的话题，发表自己的看法和评论，这些文本信息通常包含作者的主观态度和情绪[1].微博文本的情感挖掘工作，为推动电子商务发展，组织民意调查，网络舆情监控等提供重要的决策依据，具有重要的科学意义和实际应用价值[2].

细粒度情感挖掘工作包括：评价对象识别，抽取情感的表达、观点持有者等[3].这是许多自然语言处理任务基础性工作，如具有观点导向的问答系统、观点摘要的生成等[4].目前，已有的情感表达抽取工作主要集中在新闻、产品评论或博客等[4-6]，其任务是识别文本中包含作者态度或者情绪的表达部分，包括两种类型：一种是直接情感表达，如“《捉妖记》太好看了，好喜欢里面的小妖王”这句话中“太好看了”、“好喜欢”；另一种是隐含情感表达，如“没有语言能力”.

在之前的工作中，情感表达抽取通常被看作是序列标注任务[4,6].这类方法是将句子看作词的序列，并且识别每一个词(项)对应的序列类标，来抽取情感部分.CRF[4]及其变形模型[6]被成功应用于这项任务中，取得了很好的效果.文献[6]提出的基于SemiCRF方法在短语级别实现抽取，与文献[4]提出的基于词语级别的抽取方法相比，获得了更好的效果.

本文研究工作是识别中文微博文本中的情感表达.微博文本有其自身的特点，如文本内容形式多样，具有不规范、口语化等特点，并且随时都有大量的网络新词和表达产生，这对情感挖掘工作提出了新的挑战[1].微博文本的上述特性会影响自然语言处理基本任务的效果，如分词和词性标注等；而基于CRF模型[4]的情感表达抽取方法往往需要依赖大量词汇及语法特征，因而在处理微博文本的抽取任务时存在不足.

有相关文献利用微博文本的这些特点来提高一些自然语言处理任务效果，如文献[7]提出利用探测微博文本非规范化形式来提高对命名实体的识别.中文微博中存在许多网络新词和不规范词，如“坑爹”、“口耐”等，本文发现它们大部分包含在句子的情感表达中.从这一现象出发，在现有工作[4]的基础上，本文提出利用对微博文本中这些新词的探测来提高文本的情感表达抽取效果，改进原有方法在处理微博文本时的不足.

常见的相关工作利用一些语言学特征和统计量来识别新词，如文献[8]提出基于词汇模式来识别新词；文献[9]提出利用词性的组合方式来探测候选新词；文献[10]提出了利用互信息、左右信息熵等来识别新词.文献[11]提出基于CRF的新词发现方法，将新词识别任务看作序列标注问题.同样，识别微博中的新词，尤其是新的情感词也是目前非常重要的研究内容[8,12].

本文在文献[4]的基础上，将新词发现融入到微博情感表达抽取任务中，建立了基于CRF的联合模型，旨在利用新词的信息来提高情感表达识别的效果.

## 1 研究动机

中文微博的文本形式多样，其中有许多不规范词和新词，如“炒鸡给力”、“坑爹”等.微博用户使用这些新词和不规范词，往往是为了表达一定的情感倾向，如“胡巴很口耐!”表达出喜欢的情感倾向；“小哇演技棒棒哒～”表达出赞美的情感倾向.受这一现象的启发，本文考虑利用对这些词的探测来提高微博文本情感表达抽取的效果.

### 1.1 微博新词

新词是指中文文本经过分词之后，不在已有词典中的词项，也称未登录词[13-14].在中文微博中，新词和不规范词的出现具有实时性，每段时间都会有许多网络新词和表达产生，并被用户大量使用.目前，常用的分词工具对于微博文本处理的效果并不理想，中文微博新词发现也是重要的研究课题.

针对本文研究的任务，除了像“坑爹”、“爆棚”等这样的网络新词外，将使用度高的不规范词，如“杯具”、“炒鸡”、“内牛满面”等也看作新词；同时，本文也把一些出现频率较高的口语化表达作为新词，如“萌萌哒”、“棒棒哒”等等.

### 1.2 微博新词的分布情况

为了说明本文的研究动机，对微博新词的分布情况进行了统计.抓取新浪微博电影热议榜下2015年9月份的微博数据，其中抽取2 000条微博内容作为测试集，并标注文本中的情感表达和新词.

此外，本文也研究开放领域的相关任务，选取了NLPCC2014会议评测任务——“开放领域中文微博的情绪表达抽取”子任务的评测数据.该公开的语

料测试集中包含 2 000 条微博，4 261 条观点句子，并且标注了句子的情感表达；本文手工标注该数据集中的新词. NLPCC2014 语料测试集中，一共标注了 69 个新词，共出现 439 次；新浪微博电影语料测试集中，共标注了 83 个新词，累计出现 631 次.

两种语料的测试集中新词的分布情况如图 1 所示，NLPCC2014 语料的新词集合中，大约 65%的新词都在文本的情感表达部分，新浪微博电影语料中的这一比例为 78%，这说明新词更倾向于分布在文本的情感表达部分，也印证了用户使用新词往往是为了表达一定的情感倾向性. 同时，也可以看出，NLPCC2014 评测语料有 9%的句子在情感表达部分包含了新词，新浪微博电影语料中的这一比例为 12%，说明对于微博情感表达抽取任务，新词是一种可以有效利用的特征来源，图 2 是情感表达和新词联合条例标准的例子. 下面将建立融入新词发现的微博情感表达抽取模型，并分析新词对其的影响.

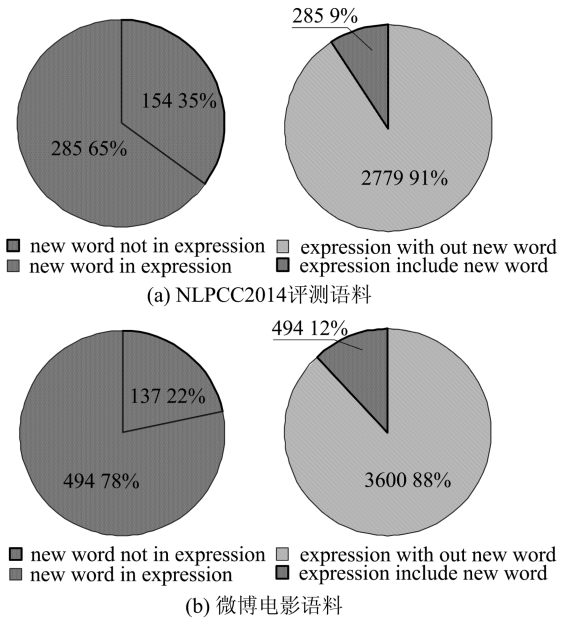


**图 1　语料中新词的分布情况**

**Fig. 1　The distribution of new words in the corpus**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 《捉妖记》 | 炒 | 鸡 | 好看 | ， | 萌 | 萌 | 哒 |
| *O* | *B*-word-*b* | *I*-word-*e* | *I*-word-*s* | *O* | *B*-word-*b* | *I*-word-*m* | *I*-word-*e* |
| 的 | 胡巴 | 真 | 口 | 耐 | ！ | | |
| *O* | *O* | *B*-word-*s* | *I*-word-*b* | *I*-word-*e* | *O* | | |

**图 2　情感表达和新词联合序列标注实例**

**Fig. 2　An example of joint model sequence annotation of emotional expression and new words**

## 2　CRF 联合模型

### 2.1　条件随机场

条件随机场(CRF)是信息抽取常用的方法，它在处理序列问题时，可以有效地引入各种特征信息[15]. 情感表达抽取以及新词发现任务，都可以看作序列标注问题.

对于观察序列为 $X=\{x_1,x_2,\cdots x_n\}$，状态序列为 $Y=\{y_1,y_2,\cdots y_n\}$ 的线性链的马尔科夫场，在给定 $X$ 和 $Y$ 序列的条件下，CRF 模型可以表示为

$$p(y\mid x;\omega)=\frac{1}{Z(x,\omega)}\exp\sum_{i=1}^{N}\sum_{j}\omega_j f_j(y_{i-1},y_i,x,i) \tag{1}$$

$$Z(x,\omega)=\sum_{y}\exp\sum_{i=1}^{N}\sum_{j}\omega_j f_j(y_{i-1},y_i,x,i) \tag{2}$$

式中，$\omega$ 为待估计参数；$f_j$ 为特征函数；$i$ 表示对输入序列 $x$ 在特征函数 $f_j$ 上的求和，这样可以保证对于变长的输入 $f_j$ 有估计 $j$ 数目的特征函数值. CRF 的一个优点在于，不用假设输入 $x$ 之间的独立性关系就能计算 $p(y\mid x;\omega)$. 输入与输出之间的关系是通过 CRF 的使用者在特定的任务中指定的特征函数 $f_j$ 以及 CRF 自动学习的参数 $\omega_j$ 来体现，情感表达以及新词的识别需要依赖相关的特征来处理. 所以本文采用线性链 CRF 模型对二者进行抽取.

### 2.2　情感表达和新词联合抽取模型

情感表达抽取任务可以看作序列标注问题，一般将序列类标集设置为：{$B$, $I$,$O$}[4,6]；其中 $B$ 表示句子中情感表达部分开始的词项，$I$ 表示情感其他词项，$O$ 表示非情感表达部分的词项. 所以，对于一个有情感倾向的句子序列，序列类标为 $B$、$I$ 的词项部分即为情感表达. 由于常用的 NLP 工具在处理微博文本时效果不佳，如分词工具不能准确识别出网络新词和不规范词而将其分割成多个词项，词性标注不够准确等，基于特征信息的 CRF 方法[4]在进行情感表达抽取时尚有一定的不足.

同样，新词发现也可以看作一个序列标注问题，

本文借鉴文献[11]中的方法，将新词相关的序列类标集合设定为{$s$, $b$, $m$, $e$}. 其中词项标签 $s$ 表示当前词项是一个已登录词；$b$ 表示新词的开始词项；$e$ 表示新词的结束词项，即新词的边界 . 如果新词词项数大于 2，即中间还有其他词项，用 $m$ 表示. 文献[11]通过 CRF 模型来识别句子的序列类标，发现了其中的新词.

本文建立了联合抽取的序列模型，如图 3 所示. 在抽取微博文本的情感表达，同时也要识别情感表达中的新词，旨在利用新词的信息. 这里将联合模型的序列类标集设置为：{$B$-word-$b$, $B$-word-oov, $I$-word-$b$, $I$-word-$m$, $I$-word-$e$, $I$-word-oov, $O$}. 其中 $B$-表示句子中情感表达部分开始的词项，$I$-表示情感其他词项，$O$ 表示非情感表达部分的词项，word-$b$ 表示新词的开始项，word-$e$ 表示新词的结尾项，word-$m$ 表示新词的中间部分，word-$s$ 表示非新词部分的登陆词.

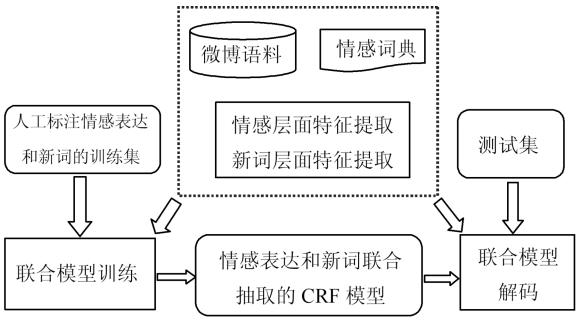


**图 3 联合抽取系统的框架**

**Fig. 3 Framework of joint extraction system**

对于上述序列模型，本文采用 CRF 算法求解，该算法是基于状态特征及转移特征来计算给出观察序列条件下状态序列类标的概率 $p(y \mid x;\omega)$，是一种判别模型. 对于训练好的 CRF 模型，如果句子中当前词项有较大的概率被识别为新词的某个状态类标，如口(-word-$b$)、耐(-word-$e$)，根据之前的对语料的分析和假设，当前词项及上下文也有较大的概率被识别为情感的状态类标.

## 2.3 特征介绍

### 2.3.1 情感表达抽取特征

对于情感特征的选取，本文参考了文献[4]中的特征类型，其中词典特征引入了台湾大学公开的情感词典作为参考，特征集合如表 1 所示.

**表 1 情感表达特征模板**

**Tab. 1 Features of emotional expression**

| 类型 | 特征 $t_i$ |
|---|---|
| 词串特征 | 前 4 个词 $\mathrm{lex}_{-4}$, $\mathrm{lex}_{-3}$, $\mathrm{lex}_{-2}$, $\mathrm{lex}_{-1}$ |
| | 当前词 $\mathrm{lex}_0$ |
| | 后 4 个词 $\mathrm{lex}_1$, $\mathrm{lex}_2$, $\mathrm{lex}_3$, $\mathrm{lex}_4$ |
| 语法特征 | 前一个词、当前词及后一个词的词性 |
| 字典特征 | 当前词在不在情感词典内 |
| 词距离 | 当前词距离情感词的距离 |

### 2.3.2 新词发现特征

新词发现特征选择是一个非常重要的步骤，本文参考文献[8,11]，选用的新词发现特征如表 2 所示，有互信息(MI)、左信息熵、右信息熵、共现频率等.

**表 2 新词发现特征模板**

**Tab. 2 Features of new word detection**

| 类型 | 特征 $t_i$ |
|---|---|
| 与相邻位置统计量 | 本词的左信息熵 $\mathrm{LE}_0$ |
| | 本词的右信息熵 $\mathrm{RE}_0$ |
| | 与前一个词的互信息 $\mathrm{MI}_L$ |
| | 与后一个词的互信息 $\mathrm{MI}_R$ |
| | 本词与前一词共同出现频率 $\mathrm{IFA}_L$ |
| | 本词与后一词共同出现频率 $\mathrm{IFA}_R$ |
| 位置信息 | 前一个词长度 $L_L$ 后一个词长度 $L_R$ |
| | 本词长度 $L_0$ |
| | 前一个词、后一个词和当前词词串和词性 |

(Ⅰ)互信息(MI) 定义为

$$M(w_1,w_2)=\frac{p(w_1,w_2)}{p(w_1)\cdot p(w_2)} \tag{3}$$

式中，$p(w_1,w_2)$ 代表词项 $w_1$，$w_2$ 共现的概率，$p(w)$为词项 $w$ 出现的概率. 互信息越大，说明两个词之间的关联性越强[10]. 本文取当前词分别与前一个词和后一个词的互信息作为 CRF 特征项.

(Ⅱ)左信息熵(LE)定义为

$$\mathrm{LE}(w)=-\frac{1}{n}\sum_{a\in A}C(a,w)\log\frac{C(a,w)}{n} \tag{4}$$

式中，$w$ 代表当前词，$A$ 代表语料中 $w$ 左边词的集合，$C(a,w)$表示 $a$ 与 $w$ 共同出现的次数，$n$ 表示 $w$ 在语料中出现的次数.

(Ⅲ)右信息熵(RE)定义为

$$\mathrm{RE}(w) = -\frac{1}{n}\sum_{a\in B} C(a,w)\log\frac{C(a,w)}{n} \quad (5)$$

式中，$w$ 代表当前词，$B$ 代表语料中 $w$ 右边词的集合，$C(a,w)$表示 $a$ 与 $w$ 共同出现的次数，$n$ 表示 $w$ 在语料中出现的次数. 词的左、右信息熵是分别衡量语料中作为词项与其左、右侧邻近词语的凝固度，熵值越大，凝固度越高[8].

## 3　实验

### 3.1　数据集和实验环境

本文第一种实验数据来源于新浪微博，抓取了2015 年 9 月份电影热议榜下的 27 000 条微博，人工标注标文本中的情感表达和新词. 选取其中的 963 条微博情感句子，包含 1 340 个情感表达，188 个新词，作为训练集；2 214 条微博句子，包含 4 094 个情感表达，631 个新词，作为测试集；抓取的所有微博条目作为提取新词统计量特征的语料.

本文还选取了 NLPCC2014 会议任务 1 提供的评测数据作为开放领域实验语料，其中情绪表达抽取子任务的训练集，包含 1 000 条微博，1 997 条情绪句子，标注了 1 639 个情感表达和 150 个新词；测试集包含 2 000 条微博，3 248 条情绪句子，标注了 3 604个情感表达和 439 个新词. 为了提取新词的统计量特征，引入了 NLPCC2014 会议情绪分类子任务的训练集，包含 14 000 条微博，45 421 条情绪句子.

针对研究的任务，本文给出了对微博文本中新词的标注方法. 首先，对于分词后的语料，采用文献[10]中计算候选新词的方法，得到大量的候选新词，新词筛选采取人工评判方法. 这里参考 COAE2014 会议任务 3 中给出的微博词典，保证本文标注的新词不包含在给定的词典内.

实验中算法采用 CRF＋＋工具包①，特征模板设置采用上文表 1 和表 2 的特征项；分词和词性标注采用文献[16]提供的开源的分词工具 ICTCLAS.

### 3.2　评价指标

3.2.1　情感表达抽取评价

对于情感表达抽取任务，采用 NLPCC2014 会议中任务 1 的评价方法，对于每个句子 $i$，精确率及召回率的定义如下：

$$\mathrm{Precision}_i = \frac{\#\,\mathrm{system\_correct}}{\#\,\mathrm{system\_proposed}} \quad (6)$$

$$\mathrm{Recall}_i = \frac{\#\,\mathrm{system\_correct}}{\#\,\mathrm{gold}} \quad (7)$$

式中，＃gold 为句子 $i$ 中人工标注情感表达部分的个数，＃system_proposed 为本文系统得到句子 $i$ 中情感表达部分的个数，＃system_correct 为与句子 $i$ 中系统和标注精确匹配得到的正确情感表达部分个数.

测试集句子级别识别的效果的评价指标定义如下：

$$\mathrm{Precision}_{\mathrm{sen}} = \frac{\sum_i \mathrm{precision}_i}{\#\,\mathrm{system\_proposed\_sentence}} \quad (8)$$

$$\mathrm{Recall}_{\mathrm{sen}} = \frac{\sum_i \mathrm{recall}_i}{\#\,\mathrm{gold\_sentence}} \quad (9)$$

$$F-\mathrm{measure}_{\mathrm{sen}} = \frac{2\times\mathrm{Precision}_{\mathrm{sen}}\times\mathrm{Recall}_{\mathrm{sen}}}{\mathrm{Precision}_{\mathrm{sen}}+\mathrm{Recall}_{\mathrm{sen}}} \quad (10)$$

式中，＃system_proposed_sentence 为本文系统得到的包含情感表达句子的总数；＃gold_sentence 为测试集中标注的包含情感表达句子的总数. $\mathrm{Precision}_{\mathrm{sen}}$、$\mathrm{Recall}_{\mathrm{sen}}$、$F$-$\mathrm{measure}_{\mathrm{sen}}$分别为测试集中句子级别的精确率、召回率和 $F$ 值. 本文以这三个评估指标评价模型的实验效果.

3.2.2　新词识别评价

对于新词识别的评价，本文也采用精确率（$p$）、召回率（$r$）和 $F$ 值三个度量标准.

$$p = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, r = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, F = \frac{2pr}{p+r} \quad (11)$$

式中，TP 表示系统正确识别的新词个数，FP 表示系统识别错误的新词个数，FN 表示系统未能识别的新词个数.

### 3.3　对比实验

本文提出的抽取系统可以同时识别微博文本中的情感表达和新词，所以本文实验从两个方面对系统的进行评估.

对于微博文本情感表达的抽取，基准实验采用文献[4]中的方法，这种方法基于 CRF 模型，采用表 1 中的特征，序列类标设置为$\{B,I,O\}$，本文将这种方法记为 CRF*. 本文在此基础上提出改进的联合模型，记为 CRF-Joint.

为了说明本文方法对新词识别的有效性，用其对模型识别新词的结果进行评估. 这里选取文献[9]

① http://crfpp.sourceforge.net/

中的方法作为Baseline对比，文献[9]主要采用互信息和基于词性组合的规则模板来抽取新词，这与本文使用了相似的新词特征，该方法在COAE2014会议微博新词发现评测任务中获得了最好的效果，本文将文献[9]的方法记为Rule&MI.

同时，本文提出的模型是有监督算法，为了进一步评价模型对新词探测的效果，本文也与文献[11]中提出的基于CRF新词识别方法进行对比实验，该方法采用和本文模型相同的新词特征. 本文将文献[11]的方法记为CRF&Score.

### 3.4 实验结果和分析

3.4.1 情感表达抽取结果

在两个数据集上的情感表达抽取实验结果如表3所示.

表3 情感表达抽取对比实验结果

Tab.3 Experimental results of emotional expression extraction

| 方法 | NLPCC2014 | | | 电影微博 | | |
|---|---|---|---|---|---|---|
| | 精确率($p$) | 召回率($r$) | $F_1$值 | 精确率($p$) | 召回率($r$) | $F_1$值 |
| CRF* | 58.90% | 38.55% | 46.74% | 71.19% | 56.25% | 62.84% |
| CRF-Joint | 59.75% | 41.01% | 48.64% | 69.28% | 60.55% | 64.62% |

对比表3中的实验结果，加入新词发现后，在两个数据集上抽取结果的$F$值都提高了2%左右. 两种方法识别准确率比较接近，但在召回率上本文方法的结果有较大的提高. 这说明无论在开放领域还是在单一领域上，新词的探测都对情感表达的识别具有较好的指示作用，所以加入新词发现后，基于CRF的抽取系统对句子的情感表达部分识别程度有一定的提升.

3.4.2 新词识别结果

对本文方法CRF-Joint做关于新词识别的效果相关评估实验，在两个数据集上新词识别的结果如图4和图5所示.

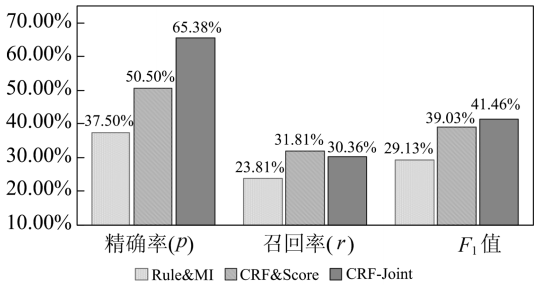


图4 NLPCC2014语料新词识别结果

Fig.4 Recognition results of new words in corpus from NLPCC2014

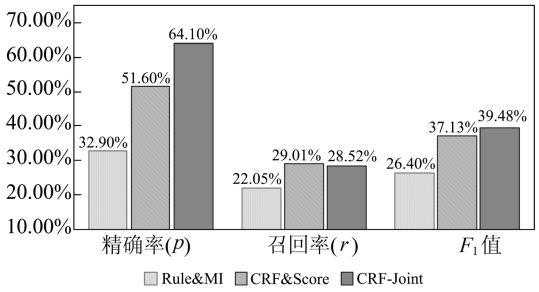


图5 新浪电影微博语料新词识别结果

Fig.5 Recognition results of new words in corpus from Sina movies' Weibo

通过与方法Rule&MI和CRF&Score对比可知，本文方法在精确率和$F$值上都有提高，说明系统对新词识别是有效的. 其中，Rule&MI采用同样的特征，但它是无监督方法，因而在识别准确率和召回率上均不如另外两种方法；CRF&Score在所有文本上识别新词，而本文提出的系统是在句子的情感部分识别新词，最终结果表明，本文方法在新词识别的精确率上有较大的提高，这说明了考虑情感新词的实际分布情况对其探测有一定的促进作用.

## 4 结论

中文微博文本的情感表达部分包含了大量的新词，本文通过构建基于CRF算法的联合抽取模型，进行对比实验论证了新词的探测对中文微博情感表达的识别具有很好的指示性. 该研究对微博情感分析工作有一定的实用价值，下一步将寻找更好的相关特征项以及探索新的联合抽取模型，进一步提高微博文本情感表达抽取的效果.

### 参考文献(References)

[1] ZHOU X, WAN X, XIAO J. Collective opinion target extraction in Chinese microblogs[J]. Chemosphere, 2013, 73(4): 532-538.

[2] 雷龙艳. 中文微博细粒度情绪识别研究[D]. 衡阳：南华大学，2014.

[3] HU M Q, LIU B. Mining opinion features in customer reviews [C]// Proceedings of the 19th National Conference on Artificial Intelligence. San Jose: AAAI Press, 2004:755-760.

[4] BRECK E, CHOI Y, CARDIE C. Identifying expressions of opinion in context[C]// Proceedings of 20th International Joint Conference on Artificial Intelligence. Hyderabad, India: Morgan Kaufmann, 2007: 2683-2688.

[5] WIEBE J, WILSON T, CARDIE C. Annotating

expressions of opinions and emotions in language[J]. Language Resources and Evaluation, 2005, 39(2): 165-210.

[6] YANG B S, CARDIE C. Extracting opinion expressions with semi-Markov conditional random fields[C]// Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Conputational Natural Language Learning. Jeju Ialand, Korea: IEEE Press, 2012: 1335-1345.

[7] LI C, LIU Y. Improving named entity recognition in tweets via detecting non-standard words [C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing: ACL Press, 2015: 929-938.

[8] HUANG M L, YE B R, WANG Y C, et al. New word detection for sentiment analysis [C]// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA: ACL Press, 2014: 531-541.

[9] 廖健,王素格,李德玉,等.基于构词规则与互信息的微博情感新词发现与判定[C]// 第二十届全国信息检索学术会议. 昆明,2014: 90-96.
LIAO JIAN, WANG SUGE, LI DEYU, et al. Using word-formation rules and mutual information for new sentiment word identification in microblogs[C]// The 20th China Conference on Information Retrieval. Kunming, 2014: 90-96.

[10] LIANG Z, XU B Y, ZHAO J. Chinese new words detection using mutual information [J]. Communications in Computer & Information Science, 2013, 320(6): 341-348.

[11] 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 Yiqun, WEI Chao, et al. Open domain new word detection using condition random field method[J]. Journal of Software, 2013, 24(5): 1051-1060.

[12] 唐波, 陈光, 王星雅, 等. 微博新词发现及情感倾向判断分析[J]. 山东大学学报(理学版), 2015, 50(1): 20-25.
TANG Bo, CHEN Guang, WANG Xingya, et al. Analysis on new word detection and sentiment orientation in Micro-blog [J]. Journal of Shandong University(Natural Science), 2015, 50(1): 20-25.

[13] LI H, HUANG C N, GAO J, et al. The use of SVM for Chinese new word identification[C]// Proceedings of the First International Joint Conference on Natural Language Processing. Berlin: Springer Press, 2005: 723-732.

[14] 李文坤,张仰森,陈若愚. 基于词内部结合度和边界自由度的新词发现[J]. 计算机应用研究, 2015, 32(8): 2302-2304.
LI Wenkun, ZHANG Yangsen, CHEN Ruoyu. New word detection based on inner combination degree and boundary freedom degree of word [J]. Application Research of Computers, 2015, 32(8): 2302-2304.

[15] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [C]// Proceedings of 18th International Conference on Machine Learning. San Francisco: Morgan Kaufman, 2001:282-289.

[16] ZHANG H P, YU H K, XIONG D Y, et al. HHMM-based Chinese lexical analyzer ICTCLAS [C]// Proceedings of the second SIGHAN workshop on Chinese language processing. Sapporo, Japan: Association for Computational Linguistics, 2003: 758-759.