一种网络非规范汉语词汇的识别方法*

姚天昉　张霄凯

上海交通大学计算机科学与工程系 上海 200240

E-mail: yao-tf@cs.sjtu.edu.cn　imshaka@yahoo.cn

摘　要：随着即时通信软件的普及，网络非规范词汇广泛出现在各种主观性文本中。在传统的文本挖掘中，这种非规范词汇都被视为噪音。事实上，这些非规范词汇经常存在于用户表达个人意愿的句子中。如果我们能够正确识别这类词汇，就能为意见挖掘提供新的意见元素信息。本文的工作把来自网络的非规范汉语词汇分为典型非规范汉语词汇和歧义非规范汉语词汇。对于典型非规范汉语词汇，我们采用了基于序列覆盖算法的模式匹配方法对其进行识别。而对于歧义非规范汉语词汇，我们则采用了基于特征抽取的分类方法进行识别。实验结果表明：上述两种方法对于识别网络非规范汉语词汇是可行和有效的。

关键词：非规范汉语词汇，网络评论，预处理，意见挖掘

An Identification Approach for Network Informal Chinese Words

Yao Tianfang　Zhang Xiaokai

Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240

E-mail: yao-tf@cs.sjtu.edu.cn　imshaka@yahoo.cn

Abstract: With the popularity of instant messaging software, there widely exist network informal words in a variety of subjective texts. In traditional text mining, such words are considered to be noises. But in fact, these informal words are often found in the sentences including individual desire expressed by users. If these words can be identified correctly, they can provide the new information of opinion elements for opinion mining. In this paper, Informal Chinese Word (ICW) from networks is divided into two types, that is, typical ICW (TICW) and fuzzy ICW (FICW). For the first one, we use sequence-covering approach based on pattern matching to identify it; while for the second one, we use classification approach based on feather extraction to recognize it. The experimental results have shown that the above-mentioned two approaches for the identification of ICWs from networks are feasible and effective.

Keywords: Informal Chinese Words, Web Review, Preprocessing, Opinion Mining.

1　概述

近年来，对描述非事实的主观性文本进行处理的研究十分活跃，这类文本的主要特点是基于断言（allegations）或评论（arguments），其内容包含有个人、群体、组织等的意见、情感和态度等。它们是文本意见挖掘研究领域的基本语料[1]，其中也包括汉语文本[2]。

* 国家自然科学基金项目（60773087）

随着互联网和个人电脑的普及，网上交流以它便捷和快速的优点，已经成为意见挖掘方面一个主要的语料来源。由于信息交流的频繁，在大多数 BBS 和网络聊天中都会存在一些不规范的语言表达形式。其特点是在规范词词典中并未出现这种表达形式，或者和规范词词典中对应词汇所表达的意思大相庭径。于是就出现了所谓的网络非规范词汇（Network Informal Word）。例如："这款车型还是受广大车迷稀饭滴！"这里的"稀饭"表达的是"喜欢"的意思。如果我们不知道"稀饭"和"喜欢"之间的对应关系，我们就无法了解这句话中意见持有者对这款车型所表达的褒义观点。所以，对网络非规范词汇的研究具有现实意义。

已有国外研究者进行过英语非规范词汇的研究工作[3]。由于英语几乎全部是字母组合的形式，所以这方面的研究工作一般都采用模式匹配方法来解决。此外，香港中文大学的研究者曾进行过繁写体汉语非规范词汇的研究工作[4][5]。这对我们的研究工作起到一定的借鉴作用。但他们的方法也存在一定的局限性：首先，他们的语料主要是基于香港地区所使用的非规范词汇，带有明显的地区特点，而且这些词汇存在许多外来词的表达（如英文缩写）；其次，他们所收集的语料规模很大（约 10 万个非规范句子），并且需要进行人工标注。由于我们课题组的研究人员有限，无法完成类似庞大的标注工程。考虑到这种限制，我们提出了一种应用最小监督的方法来处理非规范的词汇。从而能达到识别这类非规范词汇，并将结果用于处理主观性文本的目的。

2 非规范汉语词词典的构造

为了获取具有一定规模的数据集，我们首先通过网络爬虫（Crawler）获取百度贴吧热门板块的网页，其时间跨度从 2007 年 6 月到 2008 年 6 月。然后从中人工筛选出非规范汉语词汇以建立非规范汉语词典 ICW_Dictionary。接着利用这个词典抓取实验所需要的训练和测试数据集。

ICW_Dictionary 词典主要以属性对（v_i，v_j）的形式来表示。其中 v_i 表示非规范词汇表达的形式（ICW），v_j 表示非规范词汇 v_i 所对应的规范词汇（FCW）。建立词典的目的是为了方便我们以后对非规范词汇进行正规化处理工作。其存储形式如表 1 所示。

No.	ICW	FCW
00003	稀饭	喜欢
00361	8	不
……	……	……

表 1 非规范词词典的存储形式

3 汉语非规范词汇的形式和类别

非规范词汇主要来源于用户的自由输入。用户为了简单快捷，就用同音或近音、字母缩写等形式来代替自己所要表达的词汇。从而形成了这种非规范的表达方式。另一方面，由于输入法所存在的缺陷以及用户的发音不准，导致用户所表达的词汇意思和规范的词汇意思有所出入，于是就会经常出现很多谐音词。以上原因造成了所谓的"网络非规范语言（NIL）"[5]。

经过筛选，我们发现非规范词汇呈现极大的不规范特性。种类繁多，形态各异。经过反复细致地分析和比较，我们将所需要处理的非规范词汇划分为以下六个大的类别（注：文中字母不区分大小写），如表 2 所示。从表 2 可以看出：如果我们针对这六个大类分别处理，会显得重复和累赘，并且不同类别中可能会有交叉的情况，那样就无法将相同的方法进行复用。于是我们考虑将整个非规范汉语词汇集合分为下面两种类型：

典型非规范汉语词汇(Typical Informal Chinese Word, TICW)：指那些包含字母、数字以及混合缩写形式代表规范语言表达意思的词汇。这种词汇的特征就是形式上很不规范。在正常的文本中，例如杂志，报刊上一般不可能出现这种词汇。

歧义非规范汉语词汇（Fuzzy Informal Chinese Word, FICW）：指那些从字面上看是规范形式的词汇，但事实上在该词汇出现的上下文中表达的却是其对应的非规范词汇的意思。在这种情况下，我们也把这类词汇看作是非规范的词汇。这种表达方式的特点就是不仅在网络上会出现这种类型的词汇，正规书面语中也经常会出现这种词汇。从这种角度来看，正规语言的含义就具有歧义了。

词汇的表达形式	该形式表达的个数	示例
拼音或英文单词首字母缩写	127	“GF”=“女朋友”
中文谐音字	36	“斑竹”=“版主”
音译字和外来表达	38	“粉丝”=“歌迷，影迷”
汉语词汇和短语	148	“白骨精”=“白领+骨干+精英”
数字	39	“94”=“就是”;“88”=“拜拜”
以上形式的混合	54	“4 人民”=“为人民”;“3q”=“谢谢”

表 2　非规范词汇的类别

4　典型非规范汉语词汇的识别

我们采用基于规则的序列覆盖（sequential covering）算法[6]来解决典型非规范词汇的问题。该算法是基于规则的一个分类算法。它经常被用来直接从数据中提取规则。规则基于某种评估以贪心方式增长，并可以从包含多个类的数据集中每次提取一条规则。

图 1 给出了序列覆盖算法的描述。其中，Learn-One-Rule 函数是为了提取一个分类规则，使得该规则能够覆盖训练集中的大量正例，不覆盖或仅覆盖少量的反例。然而，若搜索空间太大，要找到一个最佳的规则就需要很高的时间复杂度。于是我们采用图 2 的算法来抽取识别该非规范词汇的规则（无需事先对训练的语料进行标注，仅利用 ICW_Dictionary 词典即可）。

对每个典型非规范词汇，通过上述算法可以找到判断该词汇是不

```
序列覆盖算法：
令 E 是训练记录，A 是属性-值对的集合{(Aj, Vj)}
令 Y0 是类的有序集{y1, y2, ……, yk}
令 R={}是初始规则的列表
for 每个类 y∈Y0-{yk} do
    while 终止条件不满足 do
        r = Learn-One-Rule(E,A,y)
        从 E 中删除规则 r 覆盖的训练记录
        追加 r 到规则列表内，R = R ∨ r
    end while
end for
把默认规则{} → yk 插入到规则列表 R 的尾部
```

图 1　序列覆盖算法

```
1 训练集合记为 S，s 为 S 中的一个实例（句子），规则集合为 R，初始为 NULL。
    若 s 中该关键词为非规范，记为正例；反之，记为反例。
2 while (s∈S != NULL) do
3   if(s 为正例)
4       r = SelectRuleFrom(s);
5       R += r ;
6       DeleteSentences(r); //删除此规则能够覆盖的其它句子
7   else
8       r = SelectRuleFrom(¬ s);
9       R += r;
10      DeleteSentence(r);
11  s = s->next;
12   return R
```

图 2　典型非规范词汇正规化算法

是非规范词汇的判断规则。例如“54”在句子“这种款式的手机直接 54 好了。” 表达了“无视”的意思，根据所收集的数据可总结出“54”对应的判断规则形式为：[any words]+[54]+[not quantity]-->[无视]。通过判断规则，我们就可以识别典型非规范词汇“54”，从而利用 ICW_Dictionary 词典对其进行替换。这样，我们就达到了正规化该句子的目的。

5 歧义非规范汉语词汇的识别

我们采用基于支持向量机的分类方法根据已收集的歧义非规范汉语词汇来预测和识别出现在一个句子里面的歧义词汇是否为非规范汉语词汇。和前面收集训练数据的方法相类似，但是与处理典型非规范词汇不同的地方是同一词汇采用相同个数的规范和非规范的句子，然后将所有包含这些正反实例的句子作为一个大的训练集合。

根据我们所需分类内容的特殊性，经过大量的观察、分析和实验，得出一些比较重要的特征，这些特征可以将大多数情况下歧义非规范汉语词汇识别和区分开来。表 3 所示为所选取的特征和选取说明。

特征	选取说明
典型非规范汉语词汇	如果句子本身就含有一个已经确定的非规范表达，相当于发表言论的作者已有此嗜好，则该句中包含的歧义非规范词汇是非规范表达的可能性就大大增加了。
表达意见、建议或含有情感表达的词汇	在主观性文本的前提下，再出现某些歧义非规范的表达，则这种表达在此句中就很有可能是非规范词汇了。
第一和第二人称	非规范词汇主要来自个人的主观表达，其中包含对某些事物的评论和断言。作者为了强调自己的观点或给他人的建议就常常会在句子中使用第一或者第二人称。
不规则的标点符号	非规范词汇的使用者本身就具有对表达内容进行不规范表示的性质，同样在标点符号的使用上也会很不规范。
带有情感色彩词汇及标点符号	在含有歧义非规范词汇的句子中若存在带有表达情感的符号，则主观性句子的可能性就大大增加了。

表 3 歧义非规范汉语词汇分类特征和选取说明

由于数据的稀疏问题会造成分类效果的明显差异，因此向量空间模型的转换在这里非常重要。传统的特征向量模型（Vector based Model）[7]是目前研究者最常用的表征文档的方式，主要使用 VSM 模型。由于包含有非规范词汇的文本不属于规范的文本，因此本文所做工作不能完全采用这种向量空间模型。为了计算每个特征（词条 w）在非规范词汇判断中的强弱关系，我们用下面这个式子来计算特征项 s 所具有的权重：

$$T(s) = \lambda^{g(s')} \bullet \prod_{w \in s} t(w) \qquad (1)$$

在公式（1）中，$T(s)$表示特征s在该句子中所具有的权重。其中s'表示s特征项包含的特征个数。$g(s')$表示s'所具有的部分权重。其中$\prod_{w \in s} t(w)$是为了计算相同特征值在不同分类情况下的衰减程度。λ是一个常数，在实验过程中可以调节，以得到最佳效果。当所有的$t(w)$都为1时，权值计算公式就退化为：$T(s) = \lambda^{g(s')}$ （2）

为了反映相同的词条在不同类别中句子中都出现的情况，我们引入下面步骤和算式来计算 $t(w)$：

- 令 $X \in \mathbf{Z}$ 是任意一个句子。其中 $\mathbf{Z}$ 表示所有句子的集合。用 $X.a_1, X.a_2, \ldots X.a_k$ 表示所出现的不同歧义非规范词汇。
- 令 $C(X)$ 表示 $\mathbf{Z}$ 集合的大小，$C(X, w)$ 表示包含词条 w 的句子数目。下式表示在任意一个句子 X 中出现 w 的条件概率：$P(w|X) = \frac{C(X,w)}{C(X)}$ （3）
- 令 $P(w|X.a_1 \vee X.a_2 \vee \ldots X.a_k)$ （4）表示词条 w 只在某一个句子中出现的概率，其目的是说明这个词条只和单个句子有关，而与规范或非规范没有任何关系。

于是，下面就可以给出计算 $t(w)$ 的过程了：$t(w) = \frac{C(X,w) - P(w|X.a_1 \vee X.a_2 \vee \cdots X.a_k)}{C(X,w)}$ （5）

为了便于公式 $P(w|X.a_1 \vee X.a_2 \vee \ldots X.a_k)$ 的计算，这里采用了一个简便的近似计算公式，即用 $P(r)$ 表示 $P(w|X.a_1 \vee X.a_2 \vee \ldots X.a_k)$。设 $P(r) = p(w|X.a_1)(1 - p(w|X.a_2)) \ldots (1 - p(w|X.a_k))$ （6） 其中的 $p(w|X.a_i)$ i = 1, 2, ..., k 可以进行单独计算。对于已经识别的歧义非规范词汇，最终通过 ICW_Dictionary 词典将其进行正规化替代。

6 实验结果与分析

非规范汉语词汇的识别性能实验总共包括两个部分。第一部分为典型非规范汉语词汇实验，第二部分为歧义非规范汉语词汇实验。我们将所有搜集的网页分为训练和测试部分，它们都是基于句子为单位的，训练句子数为 5791 个，测试句子数为 3722 个。

实验评价还是应用类似文本分类的评价标准，即精确率，召回率和 F 值。

$$Prec. = \frac{TP}{TP + FP} \quad Rec. = \frac{TP}{TP + FN} \quad F\text{-}Measure = \frac{2 \times Prec. \times Rec.}{Prec. + Rec.} \qquad (7)$$

上式中 TP 表示 True Positive(正确的并且选出来了)，FP 表示 False Positive(不是正确的但是选出来了)，FN 表示 False Negative(是正确的但是没选出来)。

典型非规范词汇识别性能的实验结果如表 4 所示。从实验结果可知，该方法的精确率较高，但是召回率较低。与香港中文大学的研究方法相比，精确率比它的要高，但召回率有所欠缺。他们的实验是基于 100，000 个句子的规模。所以，召回率不高说明我们搜集的典型非规范词汇还不够全面，应该需要进一步拓展语料规模。

Dataset	Precision	Recall	F-measure
TICW	0.871	0.682	0.765

表 4 典型非规范汉语词汇识别性能

在歧义非规范词汇的识别实验中，因为典型非规范词汇是歧义非规范词汇的一个重要特征。在这个实验过程中，采用了 10 折交叉验证的支持向量机分类模型。对于所有 9513 个句子作为训练和测试数据。然后针对不同的特征组合，得到了表 5 所示的实验结果。注意表 5 中特征下标的意义： a. 典型非规范词汇；b. 表达意见、建议或含有情感表达的词汇；c. 第一和第二人称；d. 不规则标点符；e. 带有情感色彩词汇及标点符号。

香港中文大学的实验结果表明：对已标注的非规范词汇的识别，最高的准确率为91.5%，最高的F值为87.1%。而从我们的实验结果来看，最高的精确率能达到92.6%，并且最高的F值为88.5%。

然而，我们所需的语料规模不到他们的十分之一。

在歧义非规范词汇的识别过程中，现在只是发现了一种可以处理的方法。但是并不是类似的这些场合都可以采用这种方法处理。在将来的工作中，要继续探索、分析非规范词汇的形成和来源。只有这样才能更好地处理非规范词汇。

特征选取	Naive Bayes			Sequential Minimal Optimization		
	Rec.	Prec.	F-Mea.	Rec.	Prec.	F-Mea.
F_a+F_b	0.531	0.662	0.589	0.476	0.711	0.570
F_a+F_c	0.708	0.677	0.692	0.635	0.783	0.701
F_a+F_d	0.675	0.813	0.738	0.832	0.806	0.819
F_a+F_c+F_e	0.734	0.810	0.770	0.794	0.765	0.779

表5 歧义非规范汉语词汇识别性能

7 结论

本文介绍了一种网络非规范汉语词汇的识别方法。由于非规范词汇是一种新型的网络语言现象，没有现成的语料供我们分析。对于已收集的非规范词汇，由于其种类繁多，且没有统一模式。在初始阶段根据其存在的形式划分为六个大类。为方便处理，我们进一步提出了将所有非规范词汇进一步划分为两个大类进行处理的方法。对于典型非规范词汇，通过基于规则序列覆盖的模式匹配算法对其进行处理。而对于歧义非规范词汇，则通过基于特征抽取的分类方法对其进行处理。

实验结果表明：同香港中文大学研究者的实验结果相比，我们的方法不需要通过大量的训练语料也能达到较为理想的效果，从而证实了所提出方法的可行性和有效性。当然，在实验以后，仍发现存在一些问题，而且这些问题是比较极端和棘手的。在今后的研究工作中，对于这些问题要进行更加细致的分析和处理，从而完善对于主观性文本中网络非规范词汇的预处理工作。

感谢：此项研究工作得到了国家自然科学基金会（项目编号：60773087）和萨尔州大学-上海交通大学语言技术联合实验室的资助。作者在此深表谢意。

参 考 文 献

[1] 姚天昉, 程希文, 徐飞玉, 汉思·乌思克尔特, 王睿. 文本意见挖掘综述. 中文信息学报, 第22卷第三期. 2008年5月. 71~80.

[2] 刘全升，姚天昉，黄高辉，刘军，宋鸿彦. 汉语意见型主观性文本类型体系的研究. 中文信息学报. 第22卷第六期. 2008年11月. 63~68.

[3] McCullagh, D. Security officials to spy on chat rooms. News provided by CNET Networks. 2004.

[4] Xia, Y., Wong K.-F., Li W. Constructing a Chinese chat text corpus with a two-stage incremental annotation approach. LREC'06. 2006.

[5] Xia, Y., Wong K.-F., Gao W. NIL is not nothing: recognition of Chinese network informal language expressions. 4th SIGHAN Workshop at IJCNLP'05. 2005. 95~102.

[6] Tan, P., Steinbach M., Kumar V. Introduction to Data Mining. Addison Wesley, 2005. 130~135.

[7] Manning C. D., Schütze H. Foundations of Statistical Natural Language Processing. The MIT Press. Springer-Verlag, 1999. 337~338.