文章编号：1003-0077（2011）00-0000-00

基于模糊推理机的汉语主观句识别*

宋洪伟，宋佳颖，付国宏

（黑龙江大学 计算机科学技术学院，黑龙江 哈尔滨 150080）

摘要：本文提出一种基于词汇模糊集合的模糊推理机以识别汉语主观句。首先，根据主、客观词概念的模糊性，我们定义了两个相应的模糊集合，并在模糊统计方法下，利用 TF-IDF 从训练语料中获取隶属度函数。然后制定了两个模糊 IF-THEN 规则，并据此实现了一个模糊推理机以识别汉语主观句。NTCIR-6 中文数据上的实验结果表明我们的方法具有一定的可行性。

关键词：主观句识别；模糊集合；模糊 IF-THEN 规则；模糊推理机

中图分类号：TP391　　**文献标识码：**A

Chinese Subjective Sentence Recognition Based on Fuzzy Inference Machine

Hongwei Song, Jiaying Song, Guohong Fu

(School of Computer Science and Technology, Heilongjiang University, Harbin 150080, China)

Abstract: This paper presents a fuzzy inference machine for Chinese subjectivity classification. To this end, we first define two fuzzy sets for lexical subjectivity and objectivity, respectively. Then, we apply TF-IDF to acquire the relevant membership functions from the training data. Finally, we define two fuzzy IF-THEN rules and thus build a fuzzy inference machine for Chinese subjective sentence recognition. We conduct two experiments on the NTCIR-6 Chinese opinion data. The experimental results demonstrate the feasibility of the proposed method.

Key words: Subjectivity recognition; fuzzy sets; fuzzy IF-THEN rules; fuzzy inference machine

1 引言

随着 WEB2.0 技术的兴起与迅猛发展，意见挖掘已经成为自然语言处理的一个研究热点[1][2]。作为意见挖掘的一个重要子任务，主观句识别的主要目的是从网络用户生成文本中将带有主观性信息的意见句从描述客观事实的客观句中识别出来。对于意见挖掘系统，主观句识别能降低系统的复杂度并提高系统的性能，因此具有极其重要的意义。

虽然近年来主观句识别相关技术已经得到快速发展，但是对于面向大规模开放性网络文本的意见挖掘系统来说，主观句识别问题仍然没有得到很好的解决。一方面，由于意见挖掘相关研究工作仍处于早期阶段，所以没有足够的标注语料用于主观句识别模型的训练；另一方面，现阶段的研究工作大部分都在概率统计的框架下看待和解决主观句识别问题，很少有人在模糊集合论的框架下，探索汉语主、客观性表示界限模糊的本质特性。因此，如果能发现主观性文本的本质特征并据此提出一种简洁的模型，对于主观句识别工作甚至是意见挖掘领域的其他工作都将具有重大的意义。

针对以上问题，本文在模糊集合论框架下，提出一种基于词汇模糊集合的模糊推理机来识别汉语主观句。首先，为了更好地识别出汉语主、客观性表示的模糊界限，我们定义了词汇的主、客观词汇模糊集合，并在模糊统计方法下，利用 TF-IDF 公式计算不同词汇分别对主、客观词汇模糊集合的隶属度。然后，本文制定了两条模糊 IF-THEN 规则，并用模糊推理

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金（60973081, 61170148）；黑龙江省人力资源和社会保障厅留学人员科技活动项目

的方法对其进行解析，以得到句子主、客观性的置信度。最后，使用重心解模糊法对得到的置信度进行解模糊操作，并利用制定的判别规则得到句子的主客观类别。

本文接下来的安排如下：第 2 节简要介绍了相关工作及背景。第 3 节描述我们方法的具体细节。第 4 节给出了在 NTCIR-6[3]中文数据上的实验结果。最后，第 5 节给出了本文工作的结论以及未来研究的展望。

2 相关研究

为了完成主观句识别任务，现阶段的研究工作大部分采用基于统计的机器学习方法来训练分类器。为了从描述性客观文本中分离出主观句，Yu 和 Hatzivassiloglou(2003)[4]提出了三个不同的方法，分别叫做基于句子相似度的方法、融合多特征的朴素贝叶斯分类器和多重朴素贝叶斯分类器。其实验结果显示多特征和多重分类器的融合对主观性识别有很大的帮助。与 Yu 和 Hatzivassiloglou 不同，Pang 和 Lee(2004)[5]将文档级别文本和句子级别文本的主观性识别任务统一起来，并形式化地将它们看作为一个面向图的最小切割的问题，由此他们实现了一个基于最小切割的主客观分类器。他们认为通过此方法，不论是传统的主观词线索还是文档内的上下文信息都能被融合起来完成主观性识别任务。蒙新泛和王厚峰(2009)[6]则研究了基于不同机器学习模型的分类器在利用上下文信息时对汉语主观句识别的影响，他们的实验表明在使用上下文信息的简单特征时，基于条件随机场模型的分类器就已经能够获得比基于支持向量机模型和最大熵模型的分类器更好的效果。近年来，为了解决标注语料稀疏问题，人们开始探索如何使用更加复杂的弱监督机器学习方法。Lin 等人(2011)[7]提出了一个叫做 subjLDA 的基于隐含狄利克雷分布的层次贝叶斯模型。与需要大规模标注语料为指导的传统分类方法不同，他们采用弱监督的生成模型学习方法，这种方法只需要少量领域相关的主观性线索词。最近，Jiang(2014)[8]则提出了一种融合多主题信息的基于隐含狄利克雷分布的弱监督机器学习模型，他的方法能够同时考虑多个主题对主观句识别任务的影响。

主观句识别任务的另一个关键问题是如何发现高质量的主观性线索。主观性特征词作为主观性线索的最小单位，最先被相关研究工作探索用来完成主观句识别的任务。首先引起人们注意的主观性特征词是形容词。Hatzivassiloglou 和 Wiebe(2000)[9]在一个简单的分类器中全面地研究了形容词的特性，包括形容词的动态极性、语义倾向及其等级对主观性识别的影响，其结果表明形容词对主观性文本具有很强的指示作用。除了形容词，Riloff 等人(2003)[10]还研究了主观性名词对主观句识别任务的影响。他们的研究表明，主观性名词虽然十分重要，但是在实际应用中很少被使用。除此之外，Wiebe 和 Mihalcea(2006)[11]的研究表明词语的词义与主观性的关联非常紧密。为了突破以单个词作为线索面临的性能上的瓶颈，一些研究工作开始尝试探索 N 元模型在主观性识别中的作用。叶强等人(2007)[12]探索基于 2-POS 模型的连续双词词类组合模式方法自动判别主观句。随后，Wilson 和 Raaijmakers(2008)[13]比较了分别用基于字的 N 元语法、词的 N 元语法和音素的 N 元语法所训练的主观性分类器的表现。除了细粒度的词汇级别线索，随后的研究工作进一步地考虑了其它粗粒度的主观性线索，比如在主观句识别任务中考虑序列模式[14]。为了自动获得大规模的序列模式，Jindal 和 Liu(2006)[15]则研究利用序列模式挖掘算法从语料中自动地提取基于类别的序列模式，进而用这些序列模式完成面向产品评论的主观性比较句识别任务。此外，Karamibekr 和 Ghorbani(2013)[16]以主观性动词为关键词，手工建立了一系列的启发式规则，进而从社会焦点评论文本中匹配出能代表主观句的主观性三元组，并以此识别主观句。

在本文中，我们处理汉语句子级别的主观性分类问题。与现存的主观性识别系统相比较，我们从模糊集合论的角度出发，提出了一种新的基于词汇模糊集合的模糊推理机来识别汉语主观句，初步的实验结果表明我们的方法能够更准确地识别出主客观句之间的细微差别。

3 主观句识别方法

在本节中，我们会详细介绍我们提出的汉语主观句识别方法，包括词汇模糊集合定义及其隶属度函数的构造方法、模糊 IF-THEN 规则和模糊推理机。

3．1 词汇模糊集合

由于自然语言本身的模糊性，词汇的主、客观性之间并没有明确的划分，这直接导致句子在主、客观性之间的模糊性。因此，本文研究利用主、客观词汇模糊集合描述词汇在主客观性之间的细微差别，进而完成汉语句子的主观性识别工作。主、客观词汇模糊集合定义如下：

定义 1 主观词汇模糊集合：设论域 X 为所有词汇的集合，则论域 X 上的主观词汇模糊集合 SUB 是 X 到 [0,1]的一个映射：

$$\mu_{SUB} : X \to [0,1] \tag{1}$$

对于 $x \in X$，μ_{SUB} 称为主观词汇模糊集合 SUB 的隶属度函数，$\mu_{SUB}(x)$ 称为 x 属于主观词汇模糊集合 SUB 的隶属度。

定义 2 客观词汇模糊集合：设论域 X 为所有词汇的集合，则论域 X 上的客观词汇模糊集合 OBJ 是 X 到 [0,1]的一个映射：

$$\mu_{OBJ} : X \to [0,1] \tag{2}$$

对于 $x \in X$，μ_{OBJ} 称为客观词汇模糊集合 OBJ 的隶属度函数，$\mu_{OBJ}(x)$ 称为 x 属于客观性词汇模糊集合 OBJ 的隶属度。

由定义可知，隶属度函数是描述模糊集合的重要组成部分，如何合理构建隶属度函数是有效应用模糊集合的关键。

3．2 隶属度函数

目前，构建隶属度函数最常见的方法有模糊统计法、参考函数法等[17]。为了避免参考函数法等方法受个人主观影响过大的缺点，本文使用模糊统计法计算每个词汇分别属于主、客观词汇模糊集合的隶属度。

模糊统计法是一种客观方法：通过 n 次重复独立统计实验来确定所有特征词中的某个特征词对主、客观词汇模糊集合的隶属度。在本文中，每次模糊统计实验主要包含以下四个要素：(1)所有特征词构成的论域 X；(2) X 中的一个固定特征词 x；(3) X 中一个随机变动的主/客观词汇集合 A^* (普通集合)；(4) X 中一个以 A^* 作为弹性疆域的主/客观词汇模糊集合 A，A 制约着对 A^* 的变动范围。

虽然模糊统计法在形式上类似于概率统计法，并且二者均是用确定性手段研究事物的不确定性。但是，模糊统计法与概率统计法分别属于两种不同的数学模型，它们有着本质区别。直观地说，概率统计方法可以理解为考察“变动的点”是否落在“不动的圈内”，而模糊统计方法则可理解为考察“变动的圈”是否覆盖住“不动的点”。

本文在模糊统计方法下利用 TF-IDF 公式构建隶属度函数。TF-IDF 公式形式简洁、实现便捷，并且相对于其他复杂的统计量，在标注语料稀疏的情况下性能更稳定，因此被广泛用于构建隶属度函数。首先，我们根据训练语料构建出一个特征词的频率矩阵：

$$A = (\mathrm{a}_{ij}), i = 1..M, j = 1..N \tag{3}$$

其中，a_{ij} 是第 i 个特征词出现在第 j 类句子中的次数，a_{ij} 指示出第 i 个特征词与第 j 类句子的关联度。M 为训练语料中的特征词个数，本文选取词频数超过 3 次的词作为特征词。N 为训练语料中句子的类别数，在本文的主观句识别任务中 N 取 2，即 1 代表主观句、2 代表客观句。

接着，为了平衡每个特征词出现在主观句与客观句中的分布，我们利用公式(4)对频率矩阵中的每个词向量进行归一化处理，经过归一化的值用 b_{ij} 表示：

$$b_{ij} = \frac{a_{ij}}{(\sum_{k=1}^{N} a_{ik}{}^{2})^{1/2}}, i = 1..M, j = 1..N \tag{4}$$

接着，我们计算了每个特征词的逆文档频率值：

$$IDF_{\mathrm{ij}} = \log_2 \frac{|D|}{|S_{ij}|}, i = 1..M, j = 1..N \tag{5}$$

其中 $|D|$ 代表训练语料中的所有句子数目，$|S_{ij}|$ 代表包含第 i 个特征词的第 j 类句子的数目。

然后，我们将公式(4)与公式(5)用乘积进行组合，以进一步地表示第 i 个特征词与第 j 类句子的关联度。此时得到的值用 c_{ij} 表示，如公式(6)所示：

$$c_{ij} = b_{ij} * IDF_{ij}, i = 1..M, j = 1..N \tag{6}$$

最后，为了满足主、客观词汇模糊集合定义中对隶属度的约束条件，我们对关联度 c_{ij} 进行归一化处理，最终得到特征词 x_i 对主、客观性词汇模糊集合 A_j 的隶属度 $\mu_{A_j}(x_i)$：

$$\mu_{A_j}(x_i) = \frac{c_{ij}}{\sum_{k=1}^{N} c_{ik}}, i = 1..M, j = 1..N \tag{7}$$

至此，我们定义了主/客观词汇模糊集合来描述主/客观词汇这两个模糊概念，并用模糊统计方法得到相应的隶属度函数。接下来，我们以上述内容为基础，在模糊推理框架下，制定和解析本文所采用的模糊 IF-THEN 规则。

3．3 模糊 TF-THEN 规则

基于模糊 IF-THEN 规则的分类模型是一种较为常见的分类方法，模糊 IF-THEN 规则被广泛的认为是分类知识较好的表示[18]。模糊 IF-THEN 规则可通过两种方法产生：自动产生方法和人工编写方法。当应用于比较复杂的系统中，基于自动产生模糊 IF-THEN 规则的方法的模糊分类系统从数据中产生规则，这样会面临大量的模糊 IF-THEN 规则，获取和优化模糊 IF-THEN 规则并不是一个很容易的任务。本文为了系统的简洁和高效，结合汉语表达的特点，选择采用人工制定的方法编写如下两条多维复合模糊 IF-THEN 规则。

R_{SUB}：IF x_1 IS 主观词汇 or . . . or x_n IS 主观词汇，THEN s IS 主观句

R_{OBJ}：IF x_1 IS 客观词汇 or . . . or x_n IS 客观词汇，THEN s IS 客观句

其中，特征词 x_i 是从训练语料中抽取得到的，n 为句子 s 所包含的特征词的数目。本文所讨论的模糊 IF-THEN 规则是一种复合模糊命题，而复合模糊命题的真值可由它所包含的原子模糊命题的真值确定。

当模糊命题 $P \in U$ 的形式为“P: x IS A”时，我们称 P 为原子模糊命题[18]。其中，x 是变量，A 是某个模糊概念对应的模糊集合。当一个模糊命题 P 是原子模糊命题时，其真值取为变量 x 对模糊集合 A 的隶属度 $\mu_A(x)$，即：$T(P) = \mu_A(x)$。

至此，本文制定了具有良好可读性和解析性的模糊 IF-THEN 规则。接下来，我们介绍如何利用模糊推理机对模糊 IF-THEN 规则进行解析。

3．4 模糊推理机

经典的推理模型本质上是一个精确的数学模型。它不仅要求规则是明确的，同时输出必须是与规则的前件相同，才能得到与后件相同的结论。当推理是从一个或几个模糊的前提推导出一个模糊的结论时，推理就成为了模糊推理，需要基于模糊数学的理论和方法来演算和处理。

针对汉语主观句识别任务，我们基于模糊数学的理论设计并实现了一个对上文提出的模糊 IF-THEN 规则进行解析的系统，本文称之为模糊推理机。模糊推理机主要有三个模块：输入模糊化模块、模糊推理模块和解模糊化模块。

(1) **输入模糊化模块**

模糊推理机的第一个阶段是对给定输入句子进行模糊化操作，即选择主观句识别系统的输入变量，并根据输入变量的隶属度函数来恰当地确定这些变量所隶属的模糊集合[17]。输入模糊化模块的具体步骤如下：

1、 首先利用查词典的方法，在输入句子 S 中找出在文档频率矩阵中出现过的特征词。

2、 然后，利用最大隶属度原则来确定特征词所隶属的模糊集合，如公式(8)、(9)所示。

$$k = \arg\max_{i \in \{SUB,OBJ\}} (\mu_i(x)) \tag{8}$$

$$\mu_k(x) = \max_{i \in \{SUB,OBJ\}} (\mu_i(x)) \tag{9}$$

其中，$\mu_k(x)$ 代表特征词 x 所具有的最大隶属度，k 代表最大隶属度对应的模糊集合。

3、 最后，对于输入句子 S，将特征词 x 划分至相应的模糊规则 R_k 的输入变量集合 A_k^S 里，而 $\mu_k(x)$ 则作为模糊规则 R_k 的实际输入值。

当得到了模糊 IF-THEN 规则的输入变量集合 A_k^S，$k \in \{SUB,OBJ\}$，我们就可以利用模糊算子对模糊规则进行推理运算。

(2) **模糊推理模块**

通常，模糊 IF-THEN 规则的前件部分具有多个输入，这时需要运用模糊算子对这些多输入进行推理，以得到一个确定数值来表示对规则后件部分的置信度。由于模糊算子是由逻辑连接词决定的，因此我们先给出本文采用的“逻辑或”基本逻辑连接词的定义。

定义 3 设 U 为模糊 IF-THEN 规则的集合，$P,Q\in U$。则 P 与 Q 的逻辑连接词“逻辑或”对应模糊集合的并运算，其真值为：

$$T(P\vee Q)=T(P)\vee T(Q)=\max\{T(P),T(Q)\}$$

显然，“逻辑或”连接词的真值与模糊集合的并运算结果是等价的。而由于模糊集合的特性，模糊集合的并运算实质上就是简单的 max 算子，这使得基于模糊集合的应用系统计算方便、可靠。但是因为客观世界现象错综复杂，简单的 max 算子已经无法适应客观世界现象赋予“逻辑或”的所有涵义。因此需要我们根据不同的任务背景，寻找合理的模糊算子以建立适合的模糊推理模型。

针对汉语主观句识别任务，本文根据模仿人脑进行模糊推理过程的特点，选择模糊集合广义并运算中的 max 算子与代数和算子。这两种模糊算子可以从不同角度解析我们的模糊 IF-THEN 规则。

通过结合原子模糊命题的真值，便可以计算出复合模糊 IF-THEN 规则的结果 τ_k^S。为了描述方便，我们将模糊算子用模糊运算符 A_t 表示，τ_k^S 与 A_t 的关系如下：

$$\tau_k^S=A_t\{\mu_k(x_1),...,\mu_k(x_i),...,\mu_k(x_n)\},x_i\in A_k^S,k\in\{SUB,OBJ\} \tag{10}$$

其中，τ_k^S 是 A_t 对句子 S 运用模糊 IF-THEN 规则 R_k 推理得到的置信度。当 τ_k^S 越大，句子 S 就越可能属于类别 *k*。

具体地，在当前句子 S 中，当模糊运算符 A_t 取上述模糊算子时，对应的形式分别为：

1、 当 A_t 为 max 算子时，$\tau_k^S=\max\limits_{x_i\in A_k^S}(\mu_k(x_i))$；

2、 当 A_t 为代数和操作时，$\tau_k^S=\sum\limits_{x_i\in A_k^S}\mu_k(x_i)$；

至此，在模糊推理框架下，我们得到了模糊 IF-THEN 规则的输出值。下一节，我们将介绍如何对模糊 IF-THEN 规则的输出值进行解模糊化。

(3) **解模糊化模块**

由于经过模糊推理后得到的是句子 S 对所有模糊 IF-THEN 规则 R_k 的置信度，因此必须进行解模糊化以将输出变为一个确定的值。常用的解模糊化方法有：重心解模糊法、最大隶属度法[17]等。本文采用重心解模糊法进行解模糊化操作，其形式如公式(11)[19]所示：

$$Y=\frac{\sum\limits_{k\in\{SUB,OBJ\}}y_k\times\tau_k^S}{\sum\limits_{k\in\{SUB,OBJ\}}\tau_k^S} \tag{11}$$

其中，y_k 是调节参数，本文通过随机梯度下降法计算其最优值。

首先，我们在最小偏差模型下，使用如下目标函数：

$$E(Y)=\frac{1}{2}\sum_{s=1}^{|S|}\left(\frac{\sum_{k\in\{SUB,OBJ\}}\tau_k^S * y_k}{\sum_{k\in\{SUB,OBJ\}}\tau_k^S}-Y'\right)^2 \tag{12}$$

然后，对其进行求偏导得到梯度函数：

$$\frac{\partial E(Y)}{\partial y_t}=\sum_{s=1}^{|S|}\frac{\tau_t^S}{\sum_{k\in\{SUB,OBJ\}}\tau_k^S}\left(\frac{\sum_{k\in\{SUB,OBJ\}}\tau_k^S * y_k}{\sum_{k\in\{SUB,OBJ\}}\tau_k^S}-Y'\right), t=1,...,N \tag{13}$$

最后，使用如下公式迭代地求解 y_k：

$$y_t(p+1)=y_t(p)-\eta * \frac{\partial E(Y)}{\partial y_t}, t=1,...,N \tag{14}$$

其中， p 为当前的迭代次数， η 为学习速率。

最终，公式(11)中的Y被映射到 $[i_{OBJ}-\Delta, i_{SUB}+\Delta]$。在本文中， i_{OBJ} 取值为0， i_{SUB} 取值1。 Δ 为系统自身的误差。为了得到识别结果，本文使用如下判别策略：

当 $Y\in[i_{OBJ}-\Delta_1, i_{OBJ}+\Delta_1]$ 时，S为客观句；

当 $Y\in[i_{SUB}-\Delta_1, i_{SUB}+\Delta_1]$ 时，S为主观句；

其中， Δ_1 为调节参数， Δ_1 的值影响系统的健壮性。为了尽可能地提高本文系统的鲁棒性，我们设定 Δ_1 取值为0.5。

至此，我们已经全面介绍了本文使用的模糊推理机的理论基础和实现细节。模糊推理机的执行过程是本文汉语主观句识别系统的核心部分，图 1 给出了模糊推理机的具体算法流程。

算法:基于模糊推理机的汉语主观句识别算法

Input : 句子 *s*

Output: 句子 *s* 的主客观类别: 主观性或者客观性

1: 预处理: 分词，词性标注;

2: *SD*(*S*) = 0

3: **for** 对 *s* 中的每个词语 *w*

4: **if** *w* 是特征词, **then**

5: 通过公式(9)计算 $\mu_k(w)$，并加入到集合 $A_k^S, k\in\{SUB,OBJ\}$ 中

6: **end if**

7: **end for**

8:通过模糊运算公式(10)计算句子 *s* 分别对主观句及客观句的置信度。

9:通过公式(11)计算句子 *s* 的模糊输出值 Y。

10: **if** $Y \in [i_{OBJ} - \Delta_1, i_{OBJ} + \Delta_1]$, **then**

11: *s* 被识别为客观句

12: **else if** $Y \in [i_{SUB} - \Delta_1, i_{SUB} + \Delta_1]$, **then**

13: *s* 被识别为主观句

14: **end if**

图 1　基于模糊推理机的汉语主观句识别算法

4 实验结果与分析

4．1 实验数据及测评方法

为了验证上述方法的有效性，我们采用 NTCIR-6[3]中文训练和测试数据，表 1 给出了数据的基本统计信息。为了评价系统的性能，本文采用 NTCIR-6 的 LWK-Lenient 评价标准给出的精确率(Precision)、召回率(Recall)和 F-值(F-score)三个评价指标。

表 1　实验数据的统计信息

项目	训练数据	测试数据
主题	4	28
文档数	143	700
句子数	2644	9246

4．2 实验结果与分析

本文第一组对比实验的目的是验证基于模糊统计 TF-IDF 方法的词汇模糊集合对主观句识别的有效性，表 2 是实验的结果。

表 2　不同特征表示法的主观句识别结果

方法	准确率	召回率	F-值
词频统计 TF-IDF	65.5%	92.9%	76.8%
模糊统计 TF-IDF	66.8%	93.7%	78.0%

在本组实验中，所使用的分类器均为基于代数和算子及重心解模糊器的模糊推理机。不同的是，词频统计 TF-IDF 方法使用传统的概率统计方法计算每个特征词的权重；而模糊统计 TF-IDF 方法则使用本文所提出的模糊统计方法来计算每个特征词的隶属度。表 2 所示的实验结果显示，在某种程度上，基于模糊统计的词汇模糊集合表示法能够更好地利用模糊推理机来区分汉语句子的主客观性之间的区别。我们分析认为，由于本文系统的出发点是希望先尽可能地区分出特征词在主客观性之间的区别，进而更准确地实现句子在主客观性之间的比较。而概率统计 TF-IDF 方法考察的是在某特定类别下所有特征词的分布情况；模糊统计 TF-IDF 方法考察的则是某特定特征词在主客观类别中的分布情况，二者之间的侧重点不同。实验结果也证明了词汇模糊集合在区分主客观的细微差别时的有效性。

为了进一步研究模糊推理机对主观性识别的有效性，本文的第二组实验对比验证了不同模糊算子对模糊推理机的影响，表 3 是实验的结果。

在本组实验中，我们采用本文提出的模糊推理机来实现主观性识别工作。为了研究不同模糊算子对模糊推理机的影响，本组实验在模糊推理阶段分别采用 max 算子及代数和算子实现“逻辑或”操作。实验结果显示，基于代数和算子的广义模糊并运算要明显好于基于 max 算子的模糊并运算，整体 F-值提高了 2.7%。我们分析认为，在基于模糊推理的主观句识别任务中，当执行“逻辑或”推理时，max 算子利用当前句子中隶属于主/客观词汇集合程度最大的特征词来代表句子从属于主/客观句的程度。这种明显的偏置性，忽视了当前句

子中的其它主/客观词汇。代数和算子则通过累加操作保留了当前句子中的所有主/客观词汇的特征，在一定程度上改善了 max 算子对高隶属度特征的偏置现象。因此代数和算子能够更好地利用模糊推理机描述汉语句子在主客观性之间的不同。

表 3　不同模糊算子的主观句识别结果

方法	准确率	召回率	F-值
max 算子	66.0%	87.7%	75.3%
代数和算子	66.8%	93.7%	78.0%

为了验证模糊推理机结合模糊集合分类模型相比于其他常用分类模型的优势，我们考察了不同分类器对模糊集合的影响。结果如表 4 所示。

表 4　不同分类方法的主观句识别结果

方法	准确率	召回率	F-值
朴素贝叶斯	65.7%	88.5%	75.4%
支持向量机	64.3%	93.0%	76.0%
模糊推理机	66.8%	93.7%	78.0%

在本组实验中，为了验证模糊推理机结合模糊集合对主观性识别的有效性，我们以词汇模糊集合为基础，研究不同类型的分类器对模糊集合的影响。实验结果显示，模糊推理机与模糊集合的组合要明显好于基于模糊集合的朴素贝叶斯分类器和支持向量机，这在一定程度上说明，相比于朴素贝叶斯分类器和支持向量机，模糊推理机能够更好地利用模糊集合来区分汉语句子的主客观性之间的区别。我们分析认为，相比于朴素贝叶斯和模糊推理机，支持向量机方法模型更加复杂，且性能容易受语料稀疏的制约。此外，虽然朴素贝叶斯分类器与模糊推理机在形式上非常相似，但是两者属于不同的模型：朴素贝叶斯模型属于生成模型，而模糊推理机是一种逻辑推理模型。由此可以看出，模糊推理机能够更好地利用模糊集合来区分汉语句子的主客观性之间的区别，实验结果也验证了模糊推理机的有效性。

表 5　本文系统与 NTCIR-6 最好系统的比较

方案	准确率	召回率	F-值
UMCP-1[3]	64.5%	97.4%	77.6%
本文系统	66.8%	93.7%	78.0%

表 5 比较了本文系统的最好结果和 NTCIR-6 中最好系统的结果。在 UMCP-1[3]系统中，他们首先采用自动获取与人工校对相结合的方法来构建情感词典，然后利用给定句子中的情感词数量判断该句是否为主观句。

实验结果显示，本文系统的最好结果较 UMCP-1[3]系统的 F-值提高了 0.4%。这在一定程度上说明，在模糊数学理论基础之上，将模糊集合与模糊推理方法有机融合具有可行性。但是在召回率方面有所下降，我们分析可能是由于训练语料稀疏使得某些特征词的隶属度估计不准确。而与 UMCP-1[3]系统相比，本文的系统可以自动地识别主观句，而无需通过手工校对的方式对情感词典进行人工维护。因此本文系统的适用性更大，能够更好地处理大规模开放性网络文本中各式各样的主观句。

5 结论与展望

本文提出了一种基于模糊推理机的汉语主/客观句分类系统，并采用 NTCIR-6 数据对系统进行了测试。实验表明我们的方法有一定的可行性，这在一定程度上说明：在模糊集合框

架下，将模糊集合与模糊推理方法融合能够很好地区分主客观句子在概念外沿上的细微区别。虽然在所进行的实验中，我们系统的准确率和 F-值达到最高，但召回率略低。我们分析可能是由于训练语料太小，这使得某些特征词的隶属度估计不准确；同时重心解模糊法的参数也得不到精确的计算。因此，在将来的工作中我们将研究如何提高特征词的质量，并进一步扩大训练语料库。

参考文献

[1] Liu B. Sentiment analysis and subjectivity[J]. Handbook of natural language processing, 2010, 2: 627-666.

[2] Pang B, Lee L. Opinion mining and sentiment analysis[J]. Foundations and trends in information retrieval, 2008, 2(1-2): 1-135.

[3] Seki Y, Evans D, Ku L, et al. Overview of opinion analysis pilot task at NTCIR-6[C]//Proceedings of NTCIR-6 Workshop Meeting. 2007: 265-278.

[4] Hong Y, Hatzivassiloglou V. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences[C]//Proceedings of EMNLP'03, 2003: 129-136.

[5] Pang B, Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts[C]//Proceedings of ACL'04, 2004: 271-278.

[6] 蒙新泛, 王厚峰. 主客观识别中的上下文因素的研究[J]. 中国计算机语言学研究前沿进展 (2007-2009), 2009: 594-599.

[7] Lin C, He Y, Everson R. Sentence Subjectivity Detection with Weakly-Supervised Learning[C]// Proceedings of IJCNLP. 2011: 1153-1161.

[8] Jiang W. Study on Identification of Subjective Sentences in Product Reviews Based on Weekly Supervised Topic Model[J]. Journal of Software, 2014, 9(7): 1952-1959.

[9] Hatzivassiloglou V,Wiebe J. Effects of adjective orientation and gradability on sentence subjectivity[C]//Proceedings of ACL'00, 2000: 299-305.

[10] Riloff E, Wiebe J, Wilson T. Learning subjective nouns using extraction pattern bootstrapping[C]//Proceedings of HLT-NAACL'03, 2003: 25-32.

[11] Wiebe J, Mihalcea R. Word sense and subjectivity[C]//Proceedings of COLING-ACL'06, 2006: 1065-1072.

[12] 叶强, 张紫琼, 罗振雄. 面向互联网评论情感分析的中文主观性自动判别方法研究[J]. 系统信息学报, 2007 1(1): 79-91.

[13] Wilson T, Raaijmakers S. Comparing word, character, and phoneme n-grams for subjective utterance recognition[C]// Proceedings of INTERSPEECH. 2008: 1614-1617.

[14] Riloff E, Wiebe J, Phillips W. Exploiting subjectivity classification to improve information extraction[C]//Proceedings of AAAI'05, 2005: 1106-1111.

[15] Jindal N, Liu B. Identifying comparative sentences in text documents[C]//Proceedings of SIGIR'06, 2006: 244-251.

[16] Karamibekr M, Ghorbani A. Sentence subjectivity analysis in social domains[C]//Proceedings of the 2013 IEEE /ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, 2013: 268-275.

[17] 张小红, 裴道武, 代建华. 模糊数学与 Rough 集理论[M]. 清华大学出版社, 2013.

[18] 阳爱民. 模糊分类模型及其集成方法[M]. 科学出版社, 2008.

[19] Rustamov S. Application of Neuro-Fuzzy Model for Text and Speech Understanding Systems[C]//Proceedings of PCI'12, 2012: 1-4.

作者简介：宋洪伟（1989—），男，研究生，主要研究领域为自然语言处理。Email: songhongwei@live.cn；宋佳颖（1990—），女，研究生，主要研究领域为自然语言处理、情感分析。Email: jy_song@outlook.com；

付国宏（1968—），男，教授，主要研究领域为自然语言处理、文本挖掘。Email: ghfu@hotmail.com。

宋洪伟： 

宋佳颖： 

付国宏（通讯作者）： 

注：1、第 1~3 位作者请随论文提供一张一寸登记照片。

2、文中图表请统一采用黑白图，文中对图的说明应与图表一一对应，表达清晰。

3、每篇论文可以注明一个通讯作者，如果需要标识，请在提交个人简介的时候标注清楚。