观点句中评价对象/属性的缺省项识别方法研究*

刘慧慧[1]，王素格[1,2]，赵策力[3]

（1.山西大学 计算机与信息技术学院，山西 太原 030006;

2.山西大学 计算智能与中文信息处理教育部重点实验室，山西 太原 030006

3.山西大学 数学科学学院，山西 太原 030006）

摘 要：在多对象、多属性的评论文本中，评价对象和评价属性的缺省识别对于观点挖掘有着重要的作用。针对情感观点句中评价对象和评价属性的缺省问题，本文提出一种有效的缺省项识别方法。首先构造缺省项识别规则集，用于获取待识别的缺省项侯选集。将缺省项识别问题看作一个二元分类问题，选用词法和依存句法作为特征，使用决策树分类算法 C4.5 训练分类器模型，在测试集上对待识别的缺省项进行判别。实验结果表明，使用依存句法特征集分类的F值优于词法特征集约2%。将词法和依存句法两类特征集融合与单类特征相比，分类精确率和F值分别提高了10%和5%左右，说明词法特征和依存句法特征的融合有利于缺省项识别。

关键词：缺省项；识别规则；词法特征；依存句法；C4.5 算法

中图分类号：TP391 **文献标识码：**A

Research on the Default Item Identification of Evaluated Object/Attribute for Opinion Sentence

LIU Huihui[1]，WANG Suge[1,2]，ZHAO Celi[3]

(1.School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China;

2.Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education Shanxi University, Taiyuan, Shanxi 030006, China

3.School of Mathematics Science, Shanxi University, Taiyuan, Shanxi 030006, China)

Abstract: The default identification of evaluated object and evaluated attribute for opinion mining is important on multi objects, multi attributes review texts. To resolve the default problem of evaluated object and evaluated attribute in sentiment opinion sentences, this paper proposes a valid identification method. At first, the rule set of default item identification was constructed for obtaining the candidate set of recognized default item. We regarded the identification of the default item as a binary classification problem, and selected the lexical and dependency parsing as features. We employed the decision tree C4.5 algorithm to train classification model which was used to judge the recognized default item on the testing date. Experimental results show that the F-value of the classification of the dependency syntactic feature set is superior to the lexical feature set about 2%. Compared with the single feature, the accuracy and F-value of the integrating of two feature sets of lexical and dependency parsing enhance almost 10% and 5% respectively, it indicates that the integrating of two feature sets of lexical and dependency parsing is helpful to the default item identification.

Key words: default item; identification rule; lexical feature; dependency syntactic; C4.5 algorithm

1 引言

微博以其短小精悍的语言特点从众多社交平台中脱颖而出，归因于它不仅是一个信息传播平台，而且是一个内容自创的平台，让人人都成为内容的制造者、见证者、传播者以及评论者。用户不仅

收稿日期：2014-06-10 **定稿日期：**2014-07-28

基金项目：国家自然科学基金资助项目(61175067，61272095)；山西省科技攻关项目(20110321027-02)；山西省回国留学人员科研项目(2013-014)；

作者简介：刘慧慧（1988—），女，硕士，主要研究方向为智能检索；王素格（1964—），女，教授，博士生导师，主要研究方向为自然语言处理、智能检索；赵策力（1986—），男，硕士，主要研究方向为智能检索。

可以发表文字内容，而且可以通过超链接、图片和视频分享资源，使得微博具有丰富的延伸性，给予用户简便的阅读体验和自由度，它要求用户发表的文字内容仅限在 140 个字数之内，因此，人们通常会使用言简意赅的语言表述对某一事物或者某一产品的看法和观点，但这导致了不规范的、口语化的文本数据日益剧增，如何从这类文本数据中挖掘所蕴含的有价值的观点，已经成为自然语言处理领域的一个热点研究课题之一[1]。

在语言表达中，人们通常省略某些语言成分，即句子存在缺省项，在相关文献中也称它为零指代[2]。它是句子中的一个缺口，指代前文中出现一个语言单位。相比于其他语言而言，汉语表达更加灵活，缺省使用也较频繁。据 Kim[3]进行调查，发现在英文文本中显式主语的使用率高达 96%，而在中文文本中显式主语的使用率只有 64%，这就意味着在中文文本中零指代的现象较为普遍。在情感观点句中，人们在不影响表达的前提下，往往使用指示性代词代替前文中所出现的某个评价对象和评价属性，或者直接将评价对象和评价属性省略。我们称前者为评价要素指代，后者为评价要素缺省。在观点要素抽取时，如果不能正确地处理评价对象与评价属性的对应关系，将导致评价对象与评价属性之间张冠李戴。例如，“苹果过于封闭，更新速度相对较慢且价格昂贵，而三星等品牌系统开放，硬件技术日益完善，手机更新速度快，受众群涵盖上、中、下三层。”，该句中评价属性“价格”、“硬件技术”对应的评价对象分别为“苹果”、“三星等品牌”。

对于评价要素指代，可以借鉴文献[4-6]中的指代消解技术。而对于评价要素缺省识别，评价对象和评价属性的缺省问题还鲜有研究。为了寻找评价对象与评价属性的关联对，需要准确识别观点句中评价对象和评价属性的对应关系，而确定缺省项的位置是其至关重要的环节。本文首先分析了观点句中评价对象和评价属性缺省项句法特点，构造候选缺省项识别规则集，在此基础上，利用句子的词性序列和候选缺省项识别规则集，获取观点句中待识别的缺省项侯选集。为了准确判定缺省项在句子中的位置，将其看作一个二分类问题。利用缺省项的上下文词性信息和依存句法信息构建分类特征集，使用决策树 C4.5 算法。在训练集上，训练分类模型，对测试集进行缺省项识别，最终获得情感观点句中评价对象或评价属性缺省项所在的位置，为实现评价对象或评价属性缺省项的恢复奠定了基础。

2 相关工作

目前，零指代识别与消解的相关研究在国内外得到了广泛地关注[2]，主要表现在以下两个方面。

基于规则方面，Kong 等[7]提出了一种基于规则探测零指代词的方法，该方法通过对一个句子进行完全句法分析，获取覆盖当前预测节点的最小子树。在此基础上，构造规则，用于确定该句子是否含有零指代词。实验结果表明，在正确的句法分析树上，F 值可达 82.45%，但在自动句法分析树上，F 值下降了近 20%。Yeh 和 Chen[8]提出了一种基于词性标注的零指代消解方法，利用一个分割程序将句子划分为带词性标注的序列，在此基础上，使用短语级解析树将其分割为更小的成分，例如名词短语和动词短语。每一个短语作为词序列，被转化为一个完整的三元组$T=[S,P,O]$。利用零指代三元组，挖掘零指代候选集，通过约束规则最终确定零指代词。实验结果表明，仅使用三元组识别零指代的精确率达到 65.2%，加上约束规则后，精确率可达到 80.5%。

基于机器学习方面，大都沿用了 Soon 等[9]提出的框架，其基本思想是将零指代消解看成二元分类问题。Ng 等[1]将零指代消解划分为零指代识别和零指代消解两个阶段，分别使用零指代词识别特征集和零指代词先行语确定特征集。在候选词选取时，他们采用了简单的启发式规则，获得大部分的零指代词，但同时也引入了太多噪音，导致前照应零指代词识别的精确率较低。Xue 等[10]给出了一种基于机器学习的空语类识别方法。该方法在完全正确的句法树上，获得了很好的结果，但在自动标注的句法树上，性能有所下降，说明句法信息对空语类识别有一定的作用。Kong 和 Zhou[11]提出了一种基于树核方法的统一框架，用于解决零指代消解问题。在零指代识别阶段，他们使用有效的句法树片段代替以往的平面特征，虽然保留了必要的上下文信息，在一定程度上提高了识别的性能，但是若句子越长，解析树越可能出现错误，并且时间复杂度也将随之增高。

对于评价对象和评价属性识别，Santosh[13]等人针对属性词抽取提出了一种无监督和领域无关的方法，整个实验过程分为三个步骤，第一步从输入的文本中识别出相关的名词短语；第二步将描述

同一个属性的名词短语聚成一类；第三步定义了属性得分函数，得分最高的侯选集即为属性词。通过在不同规模的数据集进行实验，证明了他们的算法具有较好的鲁棒性。Katharina[14]等人利用半监督学习技术抽取属性值-评价词关系对，首先自动地从未标注的数据中抽取一个初始化种子列表，将其作为半监督分类算法的训练集，最后使用依存信息和 co-location 得分建立了属性词和评价词之间的关系。

本文的研究目标是对情感观点句中缺省的评价对象和属性进行识别，通过挖掘缺省项识别规则集，选取缺省项侯选集，最后通过机器学习方法对缺省项进行识别。

3 缺省项类型

根据文献[12]，一个中文句子一般包括一个或者几个分句。依据中心理论，一个句子中，主语最可能被指代，其次是宾语，最后是其它名词。在以往的零指代研究中，侧重于处理前照应零指代，即零指代词出现在先行语之后，并且零指代词在句子中做主要的句法成分。与零指代识别研究不同，在多对象评论文本中，一个观点句可能涉及多个对象/方面。如图 1 所示。

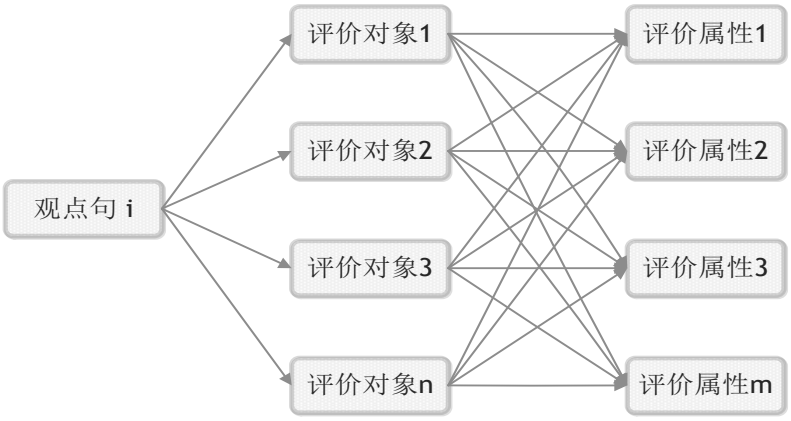

图 1 观点句—评价对象—评价属性关系对应图

图 1 中，观点句 i 可能涉及 n 个评价对象，每个评价对象可能涉及 m 个属性。

通过对大量情感观点句考察，将评价要素缺省项分为以下两种情况。

（1）缺省项作为句子的主要成分

例 1：三星太她妈难用了，还是 iphone 好，任何手机都比不上。

在例 1 中，第 3 个子句缺省了评价对象“iphone”，它作为句子的宾语。

例 2：三星手机质量太差，一进水就不好用，而且不禁摔。怀念诺基亚。

在例 2 中，第 2 和第 3 个子句中缺省了评价对象“三星手机”，它作为句子的主语。

（2）缺省项不作为句子的主要成分

例 3：新机 nexus 4 入手，外观比我想像中还要大气。手机的速度不是我吹水，真的比三星的 9300 快多了。

在例 3 中，第 2 个子句缺省了评价属性“外观”的评价对象“新机 nexus 4”。在第 4 个子句中缺省了评价对象“三星的 9300”的评价属性“手机的速度”。

4 缺省项识别框架

根据第 3 节介绍的缺省项类型，本文提出一种缺省项识别方法，框架如图 2 所示。

根据图 2，首先，初始文本经过分词和词性标注预处理，利用情感词典识别情感观点句。在此基础上，构造缺省项识别规则集获取待识别的缺省项侯选集。在训练阶段和测试阶段分别提取特征，使用决策树 C4.5 算法训练分类器模型，将其用于测试集，最后得到观点句的缺省项识别结果。

5 缺省项识别规则挖掘算法

为了获取缺省项候选集，人们通常依据语言现象总结启发式规则，但在开放的网络平台和文本大数据中，仅仅依靠人工无法将所有的情况包括其中。为了减少人为因素，我们使用缺省规则挖掘算法以其获取一个全面、科学的规则集。

定义 1：根据文献[12]，设 A 是一个由规则构成的集合，则称 A 为项集。若 A 中包含 k 个规则，则称其为 k 项集。

定义 2：设 $S=\{s_1, s_2, \cdots, s_t\}$为所有句子的集合，项集 A 在句子集 S 中出现的次数占 S 中总句子数的百分比称为项集 A 的支持度（support）。

定义 3：如果项集的支持度超过用户给定的最小支持度阈值（Min-support），则称该项集为频繁项集（或大项集）。

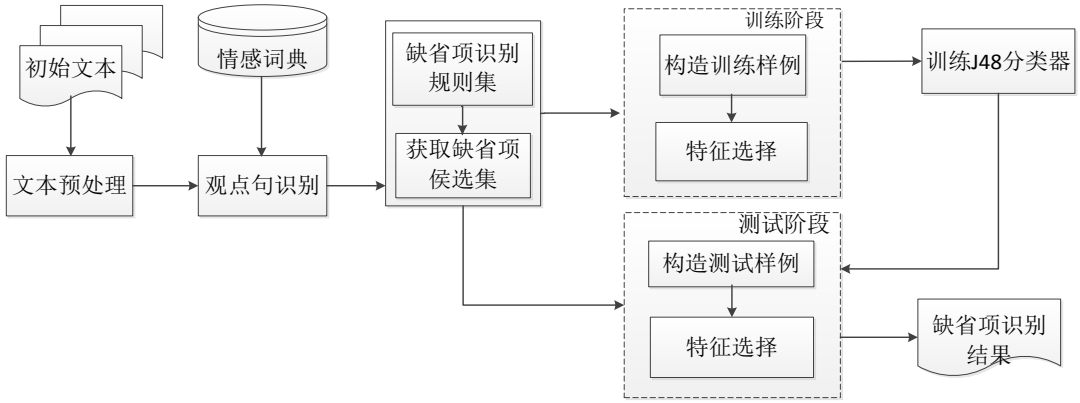

图 2 缺省项识别框架

形如规则 $X \rightarrow Y$，X 是规则的前件，Y 是结果。只有当 $X \rightarrow Y$ 的支持度和置信度分别大于最小支持度和最小置信度时，X 与 Y 之间存在关联关系。$X \rightarrow Y$ 支持度和置信度计算公式如下：

$$\text{supp}(X \rightarrow Y) = \frac{\text{同时包含X与Y的句子}}{\text{所有的句子}} \times 100\% \quad (1)$$

$$\text{conf}(X \rightarrow Y) = \frac{\text{同时包含X与Y的句子}}{\text{包含}X\text{的句子}} \times 100\% \quad (2)$$

为了获得选取缺省项侯选集的规则集，本文利用缺省项识别规则挖掘算法获取规则集，算法流程图如图3所示。

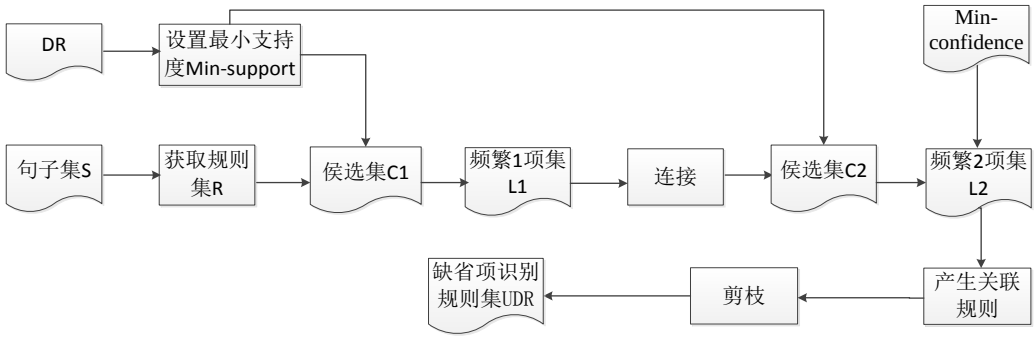

图3 缺省项识别规则挖掘算法流程图

根据图3的算法流程图，缺省项识别规则挖掘算法描述如下：

算法说明：*Li*、*Ci* 分别为频繁 i 项集和 i 项集侯选集（*i = 1，2，…，m*）；*DR* 为启发式缺省项识别规则集，它是通过对观点句缺省位置的考察，利用该位置的上下文信息，总结得到的规则集；*DF* 为 *DR* 中规则的频度集；*DAR*、*UDAR* 分别为确定性关联规则集和非确定性关联规则集。*frequence*(x)为 x 出现的次数，本文最小置信度 *Min-confidence* 设置为*0.6-1.0*，窗口大小

Window_size=i+1,i=1，2，3。

输入：序列化之后的句子集 $S=\{s_1, s_2,\ldots, s_t\}$，*DR*，*Min-confidence*，*DAR*=∅，*UDAR*=∅，*C1*=∅，*C2*=∅，*L1*=∅，*L2*=∅。

输出：缺省项识别规则集 *UDR*。

Step1: 设置最小支持度 *Min-support*

设 $DR=\{r1_k\}(k=1,2,3,...,n)$，$DF=\{f(r1_k)\}$，$Min\text{-}support=\min\{x\in DF\}$。

Step2：获取规则集 *R*

从句子 s_i 中截取 *Window_size* 长度的规则集，记为 R_{i+1}(i=1，2，3)。

Step3: 选取候选规则集 *C1*

候选规则集 $C1=C1\cup R_2\cup R_3\cup R_4$，并记录规则 r_{ij}^1 (i=2,3,4;j=1,2,…,m)的频率 $f(r_{ij}^1)$。

Step4: 产生频繁 1 项集 *L1* //频度大于最小支持度阈值的规则组成的集合。

逐一读取 C1 中的每一条规则及频率，如果 $f(r_{ij}^1)\geq Min\text{-}support$，则 $L1=L1\cup\{r_{ij}^1\}$。

Step5: 连接，即 *L1* 与自身连接

将 *L1* 的非空真子集与自身连接，产生候选 2 项集的集合，记为 *C2*。

Step6: 产生频繁 2 项集 *L2*

统计 *C2* 中两个规则 r_{ij}^1，r_{ij}^2 的支持度，如果 $f(r_{ij}^1\wedge r_{ij}^2)\geq Min\text{-}support$，则 $L2=L2\cup\{r_{ij}^1, r_{ij}^2\}$。

Step7: 由频繁项集产生关联规则

对于 *L2* 中每个非空真子集 a，如果 *frequence*(L2)/*frequence*(a)≥*Min-confidence*，则 $a\rightarrow(L2\text{-}a)$ 是一个关联规则，$UDAR=UDAR\cup\{a\rightarrow(L2\text{-}a)\}$。

Step8: *DAR* 生成

遍历 *DR* 和 *L1*，取 *DR* 中的元素 dr_i，*L1* 中的元素 $l1_j$，构造 $dr_i\rightarrow\{l1_j\}$，*i=1，2，…，*|*DR*|; *j=1,2，…，*|*L1*|的关联规则。如果 *frequence*($dr_i\wedge\{l1_j\}$)/*frequence*(dr_i)≥*Min-confidence*，则 $DAR=DAR\cup\{dr_i\rightarrow\{l1_j\}\}$。

Step9: 剪枝，*将 UDAR* 中无关的规则剔除。

遍历 *UDAR* 中每个元素 x，如果 x 前件不包含在 *DAR* 中元素的后件组成的集合中，则将其从 *UDAR* 中剔除。

Step10: 生成缺省规则集 *UDR*

UDAR 中元素的前件和后件逐一加入 *UDR* 中。

Step11: 算法结束。

6 特征选择与分类器构造

6.1 特征选择

本文将缺省项识别的过程看作一个二元分类问题，通过引入词法特征和依存句法特征，建立一个缺省项识别分类器。

（1）词法特征

缺省项位置上前后词语的词性决定了它在句子中的句法成分，而一个句子的句法成分是否完整，对缺省项识别非常关键，因此本文使用缺省项 ф 位置上前后词语的词性用于刻画缺省项的特征。

例 4：Φ 真心/d 是/v 我/r 买/v 过/u 最/d 好/a 的/u 手机/n 。/w

从例 4 中可以看出，ф 之后是副词，之前没有任何词，那么这个位置存在缺省。由此可见，词法特征可以确定缺省项的位置。

根据语料中的语言现象，词法特征描述见表 1 所示。

利用表 1 的描述，例 4 的词性特征即为 After_adv，其值为 Y。

（2）依存句法特征

虽然词法特征在一定程度上反应了缺省项的特点，但是这种平面特征只考虑了缺省项前后词的词性，往往忽略了缺省项与上下文之间的关系。为了弥补这种缺陷，我们利用依存句法分析树建立句子中词语与词语之间的联系，以其刻画词语之间的关系。

本文直接利用哈工大的依存句法树自动获取依存信息，构建 6 个依存句法特征，如图 4 所示。

表 1　词法特征集的描述

特征	特征描述
After_verb	如果 ϕ 之后是动词，则 Y；否则 N
Pre_noun_or_pron	如果 ϕ 之前是名词或者代词，则 Y；否则 N
After_noun_or_adj	如果 ϕ 之后是名词或者形容词，则 Y；否则 N
Pre_adv	如果 ϕ 之前是副词，则 Y；否则 N
Pre_verb	如果 ϕ 之前是动词，则 Y；否则 N
After_adv	如果 ϕ 之后是副词，则 Y；否则 N

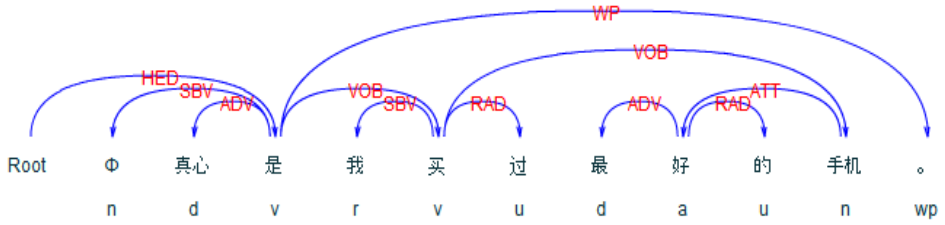

图 4　依存句法分析结果图

从图 4 中可以看出，缺省项 ϕ 与“是”之间形成了主谓关系（SBV），而且只作为从属词（箭尾），不做支配词（箭头）。

根据缺省项的上下文依存句法信息，本文构造了 5 个依存句法特征，特征集描述见表 2 所示。

表 2　依存句法特征集的描述

特征	特征描述
SBV	如果 ϕ 存在主谓关系时，则为 Y；否则为 N
VOB	如果 ϕ 存在动宾关系时，则为 Y；否则为 N
ADV	如果 ϕ 存在状中结构时，则为 Y；否则为 N
ATT	如果 ϕ 存在定中关系时，则为 Y；否则为 N
Only_father_no_child	如果 ϕ 只作为从属词，不做支配词时，则为 Y；否则为 N

依据表 2 的描述，例 4 的依存句法特征即为 SBV。

6.2 决策树分类器

决策树学习是一种临近离散值目标函数的方法，它对错误有很好的健壮性，而且适用于属性值较少的情况。本文采用决策树 C4.5 作为分类器。在训练阶段，将缺省项侯选集的每个实例通过上述表 1 和表 2 的特征集表示，对每个实例打上类标签，使用 weka 中的决策树 J48 训练分类器模型。在测试阶段，同样地，向量化每个实例，然后使用训练好的分类模型预测每个实例所属类别。

7 实验结果与分析

7.1 实验语料

本文选自 2014 年中文文本倾向性分析评测（COAE 2014）中手机领域的 292 篇微博作为实验数据，使用山西大学情感词典（共计 17445 个情感词）识别观点句，将包含情感词的句子当作情感观点句，并在情感观点句（共计 1077 个子句）上标注了缺省项的位置以及类型，如表 3 所示。该语料中共包含 848 个缺省项，其中，零指代缺省（ϕ_Z）占 45.7%，非零指代缺省（ϕ_N）占 24.3%，其他类型占 30%。

为了进一步说明仅使用情感词典判断情感句对最终实验带来的影响，本文在所有的句子（共计 1337 个子句）上标注缺省项，实验结果见表 3。

表 3　缺省项类型统计结果

编号	缺省类型	类型描述	情感句		所有句子	
			数量	比例	数量	比例
1	零指代缺省（ϕ_Z）	缺省的评价对象或者属性作为句子的主要成分	387	45.7%	422	45%
2	非零指代缺省（ϕ_N）	缺省的评价对象或者属性不作为句子的主要成分	206	24.3%	210	22.3%
3	其它	缺省的成分不是评价对象或者评价属性	255	30%	307	32.7%

由表 3 可知，仅适用情感词典判断情感句，必然会造成部分ϕ_Z和ϕ_N缺失，但相比所有句子的ϕ_Z和ϕ_N，它们在情感句中的比例略高，而第三种类型的缺省却有所上升。本文只针对前两种缺省进行处理，而使用情感词典判断情感句可以有效的减少噪音（第三种类型缺省）数据的引入。

7.2 语料校对

在手机领域中，新功能、新型号以及新别称层出不穷，由于分词软件词库未能将全部的新词收录，从而造成错分、错标等问题。为了减少预处理阶段对本文方法产生不良影响，我们对自动分词与词性标注后的评价对象和评价属性进行了校对。

（1）分词错误

例 5：三星/nz 这/r 款/q 手机/n 之所以/c 让/v 我/r 满意/v ，/w 是因为/c 自/a 拍/j 是/a 200 万/m 像/d 素/a 的/b 。/w

例 5 中的“自/a　拍/j”、“像/d　素/a”是手机的属性，应进行合并，并标注词性为“n”。

（2）词性标注错误

例 6：“9300/m 好/a 了/y ，/w 原来/d 是/v 颓废/a 的/u 包/n 的/u 问题/n”

这里“9300/m”是三星手机的一个型号，经过校对标注为“nz”。

7.3 实验结果与分析

根据第 4 节缺省项识别框架和第 7.2 小节语料校对，设计如下实验。

（1）语料校对对缺省项侯选集选取的影响

为了说明语料校正前后对实验结果的影响，我们针对缺省项侯选集 DR 方法设置了对比实验，实验结果见表 4。

表 4 语料校对前后对缺省项侯选集选取的影响

DR 方法	缺省项个数	缺省项侯选集选取的结果		
		P	R	F
语料校对前	3087	0.171	0.715	0.276
语料校对后	3088	0.179	0.750	0.289

由表 4 可知：使用相同的规则集 DR，在语料校对前后得到的缺省项的个数几乎没有发生变化，但缺省项侯选集选取的召回率有明显地改变，说明语料经过校对后在一定程度上可寻找出更多的缺省项。

（2）规则的置信度对缺省项侯选集的影响

由于规则集的大小受规则置信度高低的制约，为了识别尽可能多的缺省项，以建立较完备的缺省项侯选集，本实验选取置信度为 0.6-1.0 的规则，用于获取缺省项侯选集，实验结果见表 5 所示。

由表 5 可以看出：

①规则挖掘算法中的置信度大小对扩充启发式缺省项识别规则集有一定的影响，规则置信度越低，扩充的规则集合就越大。

②随着置信度增大，规则集的规模、缺省项个数以及规则的召回率均减小，而缺省项识别的精确率和 F 值均有增长。

表 5 规则的置信度对缺省项侯选集的影响

方法		规则集规模	缺省项个数	缺省项侯选集选取的结果		
				P	R	F
DR		73	3088	0.179	0.750	0.289
UDR 置信度	0.6	108	4244	0.155	0.894	0.265
	0.7	92	3981	0.162	0.875	0.273
	0.8	85	3641	0.168	0.830	0.280
	0.9	81	3591	0.169	0.825	0.281
	1.0	81	3591	0.169	0.825	0.281

（3）特征对缺省项识别的影响

为了验证本文构造各类特征集对缺省项识别的影响，分别考察了使用不同特征集的分类效果。与此同时，使用 Zhao[2]提出的启发式规则作为本文的 baseline。最终的实验结果采用五倍交叉验证，实验结果见表 6。

表 6 缺省项识别结果

模型		P	R	F
Baseline		0.166	0.805	0.275
本文的方法	词法特征	0.571	0.324	0.399
	依存句法特征	0.545	0.341	0.418
	词法特征+依存句法特征	0.663	0.356	0.460

从表 6 可以得知：

① 当使用依存句法特征对缺省项识别时，召回率和 F 值均略高于词法特征，从而说明依存句法涵盖的缺省项上下文信息更为丰富。

② 将词法特征和依存句法特征融合之后，精确率要远远优于任一单类特征，而融合的特征集在召回率和 F 值也有较为明显的提高，说明词法特征和依存句法特征之间具有互补性。

（4）错误分析

通过对情感观点句的评价对象和评价属性缺省项识别结果的分析，得出识别错误的主要原因有以下三个方面：

① 缺省项 Φ 的词性错标：由于在利用依存句法工具之前，人工已标记了句子的缺省项符号，导致依存句法工具对个别句子进行句法分析时，产生缺省项 Φ 的词性错标。

例如，“φ1 感觉 φ2 真不错”的依存句法图如图 5 所示：

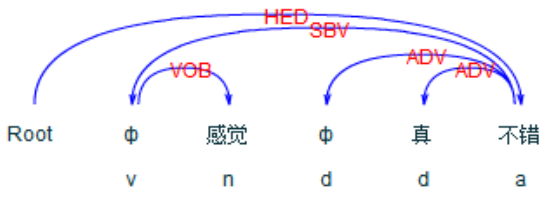

图 5 依存句法分析图

其中，“感觉”是一个动词，却被误标成了名词，导致与缺省项Φ1之间的关系发生错误，“ϕ2”的词性应该是名词，却被误标成副词，使形成的依存关系也出现错误。

② 词性标注错误：微博中的表情符号有着重要的意义，但是在分词时，往往会被冠以某一种词性，例如，“屏幕/n 大/a 又/d 4核/n ~/n ϕ3 太/d 爽/a 了/v”，其中“~”被标注为名词，使用词法特征分类时，“ϕ3 ”被认为不是缺省项，又因为“ϕ3 ”之前是名词，之后是完整的谓语，故机器误认为这个句子不存在缺省。

③ 结构化信息过少：本文主要针对两种类型的缺省项识别，一类是在句子中做主要成分的零指代缺省项，另一类是不做主要成分的非零指代缺省项。从实验结果中，可以看出句法特征 SBV、VOB、ADV 对于零指代缺省项的识别效果较好，但是对于非零指代缺省项的识别，效果不太理想，例如，“Φ 质量很差”，Φ 与“质量”之间形成 ATT 的关系，经常被错标成其他关系类型，导致非零指代的缺省项识别结果较差。

8 结束语

针对评价要素缺省项识别的问题，本文提出了一种有效的解决方法。首先使用山西大学情感词典，将包含情感词的句子作为情感句。在以往的零指代识别中，通常利用启发式规则获取侯选集，虽然简单，但也引入了过多的噪音数据，为了避免噪音数据带来的影响，本文在情感观点句上，采用缺省项识别规则挖掘算法得到规则集，用于获取缺省项侯选集。从实验结果中可以得知，使用规则挖掘算法得到的规则集优于简单启发式规则。最后，本文在缺省项候选集的基础上，构造了两类特征集用于缺省项识别的分类器，从实验结果可知，两类特征的融合要优于单类特征，从而也证明了本文方法的有效性。

本文方法的不足是整体召回率还偏低，说明构造的特征集还不够完善。未来工作中，将寻找更好的特征方法以利于缺省项识别，在此基础上，开展缺省项消解方面的研究工作。

致谢

本文使用的依存句法工具来自哈工大信息检索研究中心的中文依存句法分析工具，在此我们特别诚挚地感谢哈工大提供的语言技术平台。

参考文献

[1] C. L. Yeh and Y. C. Chen. Using zero anaphora resolution to improve text categorization. In Proceedings of the 17th Pacific Asia Conference, 2003, 423-430.

[2] S. H. Zhao and H. T. Ng. Identification and resolution of Chinese zero pronoun: a machine learning approach, In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007, 541-550.

[3] Young-Joo Kim. Subject/Object drop in the acquisition of Korean: A Cross-linguistic Comparison. Journal of East Asian Linguistics, 2000, 325-351.

[4] R. Mitkov. Robust pronoun resolution with limited knowledge. In Proceedings of the 18th International Conference on Computation Linguistics,1998, 869-875.

[5] S. Converse. 2006. Pronominal anaphora resolution in Chinese[D]. Ph.D. Thesis, University of Pennsylvania. http://www.researchgate.net/Publication

[6] G. D. Zhou, F. Kong and Q. M. Zhu. Context-sensitive convolution tree kernel for pronoun resolution. IJCNLP'2008, 25-31.

[7] K. W. Qin, F. Kong, P. F. Li and Q. M. Zhu. Chinese zero anaphor detection: rule-based approach. Advances in Intelligent and Soft Computing, 2011, 403-407.

[8] C. L. Yeh and Y. C. Chen. Zero anaphora resolution in Chinese with shallow parsing. Journal of Chinese

Language and Computing, 2007, 41-56.

[9] W. Soon, H. Ng and D. Lim. A machine learning approach to coreference resolution of noun phrase. Computational Linguistics, 2001, 521-544.

[10] Y. Q. Yang and N. W. Xue. Chasing the ghost recovering empty categories in the Chinese Tree -bank. In Proceedings of the Coling'10 Beijing, 2010,1382-1390.

[11] F. Kong and G. D. Zhou. A tree kernel-based unified framework for Chinese zero anaphora resolution, In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, 882-891.

[12] Y. Huang. Anaphora: A cross-linguistic study. Oxford, England: Oxford University Press.

[13] R. Santosh, P. Prasad, V. Vasudeva. An Unsupervised Approach to Product Attribute Extraction[C]//Proc. of the 31st European Conference on IR Research. Toulouse, France:[s.n.], 2009:796-800.

[14] P. Katharina, G. Rayid, K. Marko, F. Andrew. Semi-supervised Learning of Attribute-value Pairs from Product Descriptions[C]//Proceedings of the 20th International Joint Conference on Artificial Intelligence.[S.I.]:IEEE Press, 2007:2838-2843.