文章编号：

基于条件随机场的评价对象缺省项识别*

唐文武[1]，过弋[1,2]，徐永斌[1]，方旭[1]

（1.华东理工大学 信息科学与工程学院，上海 200237；

2.石河子大学 信息科学与技术学院，新疆 石河子 832007)

摘要：在电商网站评论文本中，评价对象和评价属性的缺省识别对文本情感分析具有重要的作用。针对电商网站评论文本中评价对象和评价属性缺省问题，本文提出了一种基于条件随机场的评价对象缺省项识别方法。首先利用情感词典识别观点句，将缺省项识别问题转换成序列标注问题，综合词法特征和依存句法特征，使用条件随机场模型进行训练，并在测试集上对待识别的观点句进行序列标注，通过标注结果判定缺省项的位置。实验结果表明，本文提出的方法具有较高的准确率和召回率，验证了该方法的有效性。

关键词：条件随机场；评价对象；缺省识别；序列标注

中图分类号：TP391　　**文献标识码：**A

The Default Item Identification of Evaluated Object Based On Condition Random Fields

Wenwu Tang[1], Yi Guo[1,2], Yongbin Xu[1], Xu Fang[1]

(1. Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai, 200237, China;

2. School of Information Science and Technology, Shihezi University, Shihezi, 832003, China)

Abstract: The default identification of evaluated objects and attributes for sentiment analysis is important in e-commerce website's review texts. To solve the default problem of evaluated objects and attributes in the review texts, this paper proposes an effective identification method based on Conditional Random Fields (CRF). At first, an emotion dictionary was applied to recognize the opinion comments. Thereafter, we treated the identification of default item as a sequence labeling problem, and selected the lexical and dependency parsing elements as features. Finally, the CRF model was trained with these features to decide the proper positions of default items. The evaluation results prove that our proposed method demonstrates its validity with reasonable good accuracy and recall rates.

Key words: Conditional Random Fields(CRF); evaluated objects; the default identification; sequence labeling

1 引言

随着互联网技术迅猛发展以及网络应用的迅速普及，互联网已经涉及到人们生活中的方方面面，并成为人们直接表达自己情感的重要平台。互联网信息的爆炸式增长，伴随着京东、天猫、亚马逊等电子商务网站发展。大量的评论是用户对商品直接情感的表达。人们在评论一个产品时，通常会使用简明的语言去表达自己的看法。因此，导致了评论文本口语化、不规范、缺省现象严重等特点。

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金（61462073）

作者简介：唐文武（1992--），男，硕士研究生，主要研究方向为自然语言处理、情感计算；过弋（1975--），男，教授，博士，主要研究方向为自然语言处理、智能信息处理、本体工程；徐永斌（1990--），男，硕士研究生，主要研究方向为自然语言处理；方旭（1993--），男，硕士研究生，主要研究方向为自然语言处理。

中文缺省也称为中文零指代[1]，是指人们在特定的语言环境下，在不影响意思表达的前提下，为了使语言简洁明快，会省去句子中的某些语言成分的现象。在情感观点句中，人们往往会省略评价对象和评价属性。评价对象是指评论所针对的对象或对象的属性。如“虾很新鲜。”，这句观点句中，“虾”作为该观点句的主语，充当该评价的对象，“很新鲜”用来修饰“虾”作为该评价对象的评价短语。

目前，在对评价元素的抽取研究工作已经取得了一定的成果，但是大多数的研究工作只能抽取出句子中存在的评价对象和属性。评价对象的缺省，导致了在进行评价要素抽取时，常常无法准确、全面得抽取出评价要素，句子中大量的评价词无法匹配到评价对象的问题。当前对于中文缺省识别的研究并不多，因此本文主要针对观点句中评价对象缺省项识别进行研究。本文将判定缺省项在句子中的位置问题转换为序列标注问题，综合词特征、词性特征和句法特征对条件随机场模型进行训练，最后利用训练后的模型识别测试集中缺省项在观点句中的位置，从而为评价对象缺省项恢复的工作奠定了基础。

2 相关研究

目前，在零指代识别问题上主要有基于规则和基于机器学习两种方法。

基于规则方面，Yeh 和 Chen 等[2]将规则方法应用到中文零指代消解的零指代项识别研究中，通过大量手工标注的规则，并提出了中心理论的方法来解决中文零指代消解。杨国庆等[3]参考 Yeh 等提出的方法，提出缺省三元规则，以动词驱动为核心提出规则来获得缺省项的结构化信息。Kong 等[4]提出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，由此获取覆盖当前预测节点的最小子树，从而构造一定的规则去判断句子中的零指代词。由于基于规则的方法主要依赖于人工构建大量的规则，将会耗费大量的人力。因此，人们更青睐于使用机器学习的方法去解决零指代问题。

Zhao 等[5]是第一个利用机器学习算法解决了零指代词识别与零指代词恢复的问题，为之后的工作提供了基础。Kong 和 Zhou[6]在同一个框架下，提出了基于树核函数的零指代识别和消解的方法，从结构化信息入手解决零指代识别问题。Song 等[7]将零指代识别和零指代消解两个子任务通过马尔科夫逻辑进行联合，在同一个机器学习框架下进行处理。秦凯伟等[8]实现了一个基于机器学习的中文缺省项识别系统，选取多个特征进行组合，利用支持向量机 SVM 进行缺省项识别研究。刘慧慧等[9]对评价对象缺省识别进行了研究，通过决策树算法对候选缺省项集进行二元分类，从而进行判定观点句中是否存在缺省现象。Yang 等[10]提出了将零指代词识别问题转换为打标签问题的方法，利用词法和语法特征，通过二元分类器为每个词打上标签，以此来识别句子中是否出现缺省现象。此外，Rao 等[11]通过模型跟踪对话中焦点的流动，对话语中的零指代问题进行了研究。Chen 等[错误!未找到引用源.]提出了一种无监督的概率模型，通过显著性模型来获取语篇信息，同时解决了零指代识别和恢复。

在目前利用机器学习进行缺省项识别的研究中，大多数都将缺省项识别转换为二元分类问题，利用标准句法信息作为特征，并在标准的句法树上获得了很好的性能，但在自动句法树上性能并不好。评价对象的缺省破坏了该对象周围正常的词串、词性和依存关系搭配序列，因此在真正的应用中获得正确的句法信息是困难的，利用标准的句法树上提取的特征训练出的模型应用在自动的句法树上导致性能的下降。由于评价对象在句子序列中出现的位置具有一定的规律性，其缺省的位置同样具有一定的规律性。通过在自动句法树上提取特征，并融合词串、词性特征，对非正常的词序列打上标签，从而可以获取评价对象缺省的位置。因此，本文提出的方法是将评价对象缺省识别转换为序列标注问题，利用依存句法树自动获取依存信息作为特征，并结合词法特征，利用条件随机场模型对评价对象缺省项位置进行识别。

3 基于 CRF 的评价对象缺省项识别

3.1 缺省项类型

在缺省项类型的分类上，许多文献都使用了 CTB[13]语料中对缺省项的分类。其分类如

表 1 所示：

表 1　CTB 中缺省项分类

类别	描述
NONE-*T*	缺省为主题或从句实施者
NONE-*	缺省在提升结构和被动结构中
NONE-*PRO*	从句中缺省明显主语
NONE-*pro*	缺省的为主语或宾语
NONE-*RNR*	发生预指的省略语形式
NONE-*?*	其他类型

其中，NONE-*T*、NONE-*PRO*以及 NONE-*pro*占的比例最大。根据以上分类的规则，以及对观点句中缺省项的观察分析，本文依据文献[9]上的分类，将观点句中评价对象缺省项的类型主要分为以下两种情况：

（1）　缺省项作为句子的主语或宾语等主要成分

例 1：虾不错，很新鲜，第二次买了。

在例 1 的第 2 个子句中，缺省了评价短语“很新鲜”的评价对象“虾”，该词作为句子的主语。

例 2：顺丰就是快，其他物流都比不上。

在例 2 中的第 2 个子句中，缺省了评价对象“顺丰”，该词作为句子的宾语。

（2）　缺省项作为非主要成分

例 3：阿根廷红虾太好吃了，价格也亲民，比白虾便宜好多。

在例 3 中的第 2 个子句缺省了属性词“价格”的评价对象“阿根廷红虾”，第 3 个子句中缺省了“白虾”的评价属性“价格”。

3.2 缺省项识别

根据中心理论[14]，主语、谓语和宾语作为句子的主要成分，其中主语是最有可能被指代，其次是宾语，最后为其他位置上的词语。因此，缺省项出现在句子中的各个位置上的概率具有明显的差异。通过机器学习的方法计算每个位置上出现缺省的概率，从而得到缺省项最有可能出现的位置。

本文将识别缺省项在情感句中出现的位置转化为序列标注问题。通过对每个词设定标签，以此判断该词之前是否出现缺省项，并利用机器学习模型解决序列标注的问题。本文将序列标注问题定义为：

定义 1 $X=(x_1,x_2,\dots,x_n)$为长度为 n 的观察序列，对于给定的观察序列，输出对应的标签序列$Y=(y_1,y_2,\dots,y_n)$，其中y_i为x_i所对应的序列标签。

在序列标注的问题上，目前有很多模型得以应用，如隐马尔科夫模型、条件随机场、自动转换机、最大熵模型以及支持向量机 SVM 等。其中隐马尔科夫模型、最大熵模型以及条件随机场是最常用最基本的三种模型，另外 SVMTool 也将 SVM 原理应用于序列标注的问题上。CRF（条件随机场）作为一种性能良好的标记和切分序列化数据的统计框架，在词性标注、命名实体识别、分词等自然语言领域都有着比较好的应用场景。CRF 在序列标注问题上克服了马尔科夫模型必须具备独立性假设的问题，可以容纳任意的上下文信息，特征设计灵活。而相比于最大熵模型，其标记偏置的缺点在 CRF 上得到了解决。考虑到上下文信息对缺省项识别的影响，以及为了能够更好得融合多个特征进行推理。因此，本文提出利用 CRF 对情感句中评价对象缺省项的位置进行识别。

在序列标注模型上，定义集合 X 为观点句中的词语，标签集合为$Y=\{N;P;O\}$；其中，N 表示该词之前存在缺省项,且作为句子的主要成分；P 表示该词之前存在缺省项，且不作为句子的主要成分；O 表示该词之前不存在缺省项。因此，利用条件随机场模型生成只包含 N、

P 和 O 的序列，则通过找到标记 N 和 P 所对应的词语，就可以判断该词之前存在缺省项。例如观点句“虾不错，很新鲜，价格便宜。”，通过 CRF 进行标注后，对应的标注序列为“虾/O 不错/O，很/N 新鲜/O ，/O 价格/P 便宜/O 。/O”，由此可知，“很”这个词对应的标签为“N”，则该观点句中评价对象缺省出现在“很”之前。

图1显示了利用CRF识别评价对象缺省项的整体流程。首先通过对评价语料进行分词、分句、清洗等预处理；然后，通过 HowNet 情感词典进行观点句的识别；接着进行特征选择、选取词法特征、词性特征和句法特征作为模型的特征；接着进行语料的标注，形成训练语料和测试语料；利用训练语料训练模型；最后利用模型进行测试语料的测试，生成缺省项识别的结果。

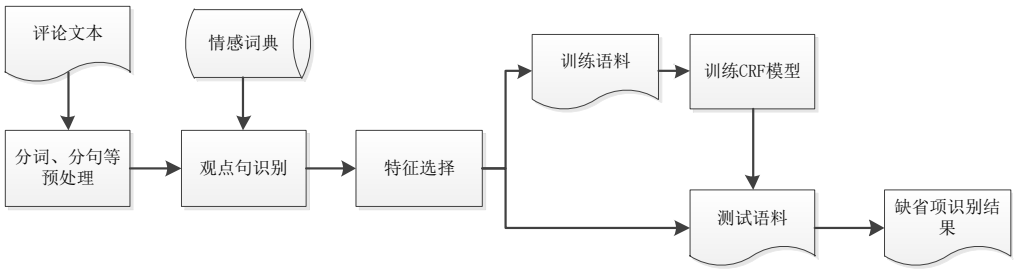


图 1 基于 CRF 的评价对象缺省项识别框架图

3.3 条件随机场模型

条件随机场模型 CRF，是由 John Lafferty 和 Andrew McCallum [错误!未找到引用源。] 在 2001 年提出的一种判别式的无向图模型，是用于切分和标记有序数据的条件概率模型。CRF 是一种性能良好的标记和切分序列化数据的统计框架模型。在词性标注、命名实体识别、分词等自然语言处理领域有着比较好的应用场景。它不仅克服了马尔科夫模型必须具备独立性假设和最大熵模型标记偏置的缺点，而且可以综合使用包括字、词以及上下文信息等多种特征，并且允许选择任意的外部特征，将特征融入到模型中。最后，在实现特征的全局归一化后，获取到全局的最优解。本文对于 CRF 做了如下定义：

定义 2 设$G(V,E)$为一个无向图，若随机变量Y_V在条件 X 的出现的情况下，其条件概率分布遵循马尔科夫特性，即满足公式 1 所示：

$$P\{Y_V|X,Y_w,w\neq v\}=P\{Y_V|X,Y_w,w\sim v\} \quad (1)$$

则称(X,Y)构成了一个条件随机场。其中，V 和 E 分别代表了无向图$G(V,E)$的顶点和边的集合，而Y_V则是 G 的顶点的索引，$w\sim v$表示在无向图 G 中w和v相邻。其模型的定义如下：

定义 3 设 X，Y 为随机变量，$X=(x_1,x_2,\dots,x_n)$为长度为 n 的待观测序列，而 $Y=(y_1,y_2,\dots,y_n)$为与 X 长度相同的状态输出序列。按照 CRF 的原理，其状态输出序列可以表示为：

$$P_\lambda^*(y|x,\lambda)=\frac{1}{Z_\lambda(x)}\exp\left(\sum_{i=1}^n\sum_j\lambda_j f_j(y_{i-1},y_i,x,i)\right) \quad (2)$$

其中，

$$Z_\lambda(x)=\sum_j\exp\left(\sum_{i=1}^n\sum_j\lambda_j f_j(y_{i-1},y_i,x,i)\right) \quad (3)$$

$$f_k(y_{i-1},y_i,x,i)=\begin{cases}t_k(y_{i-1},y_i,x,i),k=1,2,\dots,k_1\\S_l(y_i,x,i),k=K_1+l,l=1,2,\dots,k_2\end{cases} \quad (4)$$

$Z_\lambda(x)$为归一化因子，$t_k(y_{i-1},y_i,x,i)$为当前位置y_i与前一个位置y_{i-1}之间的转移特征函数；$S_l(y_i,x,i)$为当前位置的状态特征函数；λ_j是需要从训练数据中学习的参数。因此，CRF 的标注任务是在给定了一个输入序列λ之后，求出搜索概率最大的y^*，且$y^*=\text{argmax}P_\lambda^*(y|x,\lambda)$。

3.4 特征选择和语料标注

在进行缺省项识别的特征选择时，不仅需要考虑词本身的特征，句子的结构特征也对缺省项的识别具有很大的影响。在大多数的研究中，都采用了语料中已经标注的正确句法信息作为特征，但在真正的应用中获得正确的句法信息是困难的。因此本文采用了词法特征及依存句法特征，如表 2 所示。

表 2 特征说明

特征	特征代号	特征意义
词串	Token	表征当前的词串
词性	Pos	表征当前词串的词性，如名词（n）
依存句法	Dln	词语之间的依存关系

（1） 词法特征

不同位置上的缺省项，其前后词语的词串和词性也不同。由于不同位置上发生缺省的概率不同，因此不同词性的词串其前后存在缺省项的概率也不相同。例如：一个句子的第一个词为动词，该词前存在缺省项的概率比名词或者代词来的大；在“她/r 说/v 很/d 干净/a”和“她/r 说/v 虾米/n 很/d 干净/a”这两句评价句的对比中可以看出，副词前一个词为动词与副词前一个词为名词两种情况相比，前者在副词前更有可能存在缺省项。由此可知，评价对象的缺省破坏了正常的词性和词串搭配，从而存在非正常的词性和词串搭配的位置更容易出现缺省项。因此本文使用词法特征作为判定缺省项位置的特征。

（2） 依存句法特征

仅仅用词法特征进行缺省项的判定是不够的，无法利用缺省项的上下文关系。中文句子中成分的排列具有一定的规律性，例如不存在主谓关系，却存在动宾关系的句子其谓语之前很有可能存在缺省项。因此本文也使用了依存句法关系特征以此来表征词语之间的关系。

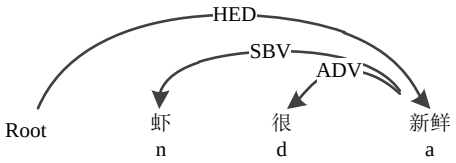


图 2 完整的句子依存关系

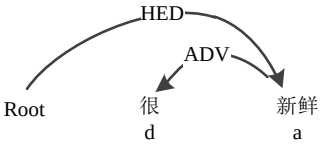


图 3 评价对象缺省的句子依存关系

在图 3 中“很”与“新鲜”存在状中结构（ADV），且“很”作为从属词（箭尾）。“新鲜”与根节点存在 HED 关系。从图 1 和图 2 的对比可以看出，存在状中关系的“很”之前存在缺省评价对象“虾”。

在语料的标注上，本文使用 3-tag 标注法。标签 N 表示当前词之前存在缺省项，且缺省项作为句子主要成分；标签 P 表示当前词之前存在缺省项，且缺省项不作为句子主要成分；标签 O 表示当前词之前不存在缺省项。在特征标注上，本文使用哈工大的自然语言处理工具 LTP，通过对情感句进行切词、词性标注、依存句法分析等对特征进行标注。其中，对于每个词的句法特征，标记为该词作为从属词时其对应的句法依存关系。训练语料和测试语料的标注样例如表 3 所示。

在表 3 中语料的标注样例中，“很”标注为 N，表示其之前存在缺省项，且作为句子的主要成分，在例句中为缺少主语。“价格”标注为 P，表示该词之前存在缺省项，且不作为句子

的主要成分，在例句中缺省了评价对象“虾”。

表 3 CRF 语料标注样例

词串	词性	句法特征	标注
虾	n	SBV	O
不错	a	HED	O
，	wp	WP	O
很	d	ADV	N
新鲜	a	COO	O
，	wp	WP	O
价格	n	SBV	P
便宜	a	COO	O
。	wp	WP	O

利用训练数据训练之后得到的 CRF 模型对测试数据进行测试，将会对每个词串进行标注，通过标注的标签得到测试数据中缺省项的位置。测试结果样例如表 4。

表 4 CRF 测试结果样例

词串	词性	句法特征	人工标注	CRF 标注
湿度	n	SBV	P	P
比	p	ADV	O	O
描述	v	POB	O	O
的	u	RAD	O	O
要	v	ADV	N	N
大	a	HED	O	O
些	q	CMP	O	O
。	wp	WP	O	O

从表 4 的结果样例中可以看出，在“湿度”之前存在缺省项，应该为“虾的湿度”，缺省了“虾”，且不作为句子的主要成分；在“要”之前同样存在缺省项，缺省了比较对象“描述的湿度”，且作为句子的主要成分。

4 实验

4.1 数据集

本文所使用的数据集是从天猫网站上采集的关于虾类商品的评论数据，抽取了其中 1980 条评论信息作为本文的语料。通过清洗、分句等预处理，最后得到 3366 条子句。在情感观点句的识别中，本文使用 HowNet 情感词典进行情感句的判断，由于考虑到词典中的词语由于词性的不同会导致情感倾向性的差异，因此在词典中加入词性信息使得情感句的判断更加准确，共识别出 2539 条观点句。在实验语料的标注上，本文采用人工标注的方法。语料中评价对象的缺省项位置的标注均由两名标注者进行手工标注，其标注结果的一致性大于 0.8，具有一定的可信度。对于语料中两人标注不一致的部分，则交由第三人进行标注。语料中评价对象缺省项类型统计结果如表 5：

表 5 缺省项类型统计结果

编号	缺省项类型	数量	比例/%
1	缺省项作为句子主要成分	1432	56.40
2	缺省项不作为句子主要成分	294	11.58
3	不包含评价对象缺省项	813	32.02

从表 5 可以看出，包含作为句子主要成分的评价对象缺省项类型的句子占所有句子总数

的 56.40%；包含不作为句子主要成分的评价对象缺省项类型的句子占所有句子总数的11.58%。因此，评价对象缺省项在本文的语料中占有 67.98%的比例。

4. 2 实验结果与分析

4.2.1 自然语言处理工具测试结果对比

本文的方法中综合了词串、词性和依存关系作为 CRF 模型的特征。在训练数据和测试数据的生成过程中，需要利用自然语言处理工具对数据进行处理。分词效果的好坏直接影响了词性和依存关系的判断。因此，为了选择合适的自然语言处理工具处理本文的短文本数据集，本文对 LTP、HANLP、FNLP 三种自然语言处理工具进行了分词测试实验。本文随机抽取了 1000 条句子分别利用三种自然语言处理工具进行了测试，并通过人工校验的方法对测试结果进行判断。测试结果见表 6。

表 6 自然语言处理工具分词实验结果

工具	切分出的总词数	切分错误数	正确率/%
LTP	4037	228	94.35
HANLP	3756	351	90.66
FNLP	3665	439	88.02

从表 6 中的实验结果可以看出，FNLP 相对于其他两种自然语言处理工具的分词结果，正确率较低，为 88.02%。其主要的错误在于对名词与形容词组合的短语往往无法进行正确的切分。例如，评价短语“质量好”中，“质量好”无法被正确切分出“质量”和“好”两个词。由于评论短文本中会出现大量类似的短语，因此 FNLP 不适合处理本文的数据。

HANLP 的分词正确率为 90.66%，其错误的最大比例在歧义的处理上。例如“活动价”则会被切分为“活动”和“价”，“快递员”则被切分为“快递”和“员”，“尝过后”会被切分为“尝”和“过后”等。LTP 的分词结果最好，正确率为 94.35%，较少出现上述两种工具的分词问题。因此在对评价数据进行处理时，本文采用了 LTP 自然语言处理工具进行处理。

4.2.2 评价对象缺省项识别实验

本实验对于测试本文提出的方法的性能，主要采用了准确率 P、召回率 R 和 F 值三种指标，其计算方法如下：

$$\text{正确率 } P = \frac{\text{算法标注为 N 和 P 的词语中正确的个数}}{\text{算法标注为 N 和 P 的所有词总数}} \quad (5)$$

$$\text{召回率 } R = \frac{\text{标注为 N 和 P 的词语中正确的个数}}{\text{语料中存在的标注为 N 和 P 的词总数}} \quad (6)$$

$$F = \frac{2}{1/R+1/P} \quad (7)$$

本实验将 2539 条观点句中，取出 2072 条观点句作为训练语料，467 条观点句作为测试语料进行实验。训练语料和测试语料的特征标注上使用了哈工大自然语言处理工具 LTP 进行处理，形成训练语料和测试语料。使用 CRF++0.53 工具进行 CRF 模型的训练以及测试。本文使用文献[9]提出的利用规则找出候选缺省项，再综合词法和句法特征利用决策树算法进行对候选缺省项判断的方法作为本文的 Baseline。另外，对不同的特征组合进行了实验，包括词串特征+词性特征、词串特征+依存语法特征、词串特征+词性特征+依存句法特征来说明特征组合对实验结果的影响。最终的实验结果见表 7。

从表 7 中可以看出，本文提出的方法相比于 Baseline 中的方法在本文语料的评价对象缺省项识别上具有明显的提高，正确率、召回率和 F 值分别为 86.03%、69.44%和 76.85%。同时，从特征组合对比实验中可以看出，综合了词法特征和句法特征后，相比于词串+词性特征和词串+句法特征的组合得到的效果更好，也验证了该方法的有效性。另外由于句子成分缺省的影响，导致在进行分词、词性标注和依存句法分析时会发生错误，这些错误也直接导

致了方法性能上的下降。

表 7 评价对象缺省项识别实验结果

模型	P/%	R/%	F/%
Baseline	67.71	36.33	47.29
词串+词性	81.90	64.85	72.83
词串+依存句法	83.86	63.82	72.48
词串+词性+依存句法	86.03	69.44	76.85

此外，本文还对香蕉商品的评论数据进行了处理。同样随机抽取了 2539 条观点句进行了实验，其中 2072 条作为训练语料，467 条作为测试语料，并通过同样的处理，最后的实验结果如表 8。

表 8 虾类和香蕉评论实验结果

商品	P/%	R/%	F/%
虾类	86.03	69.44	76.85
香蕉	84.90	71.29	77.50

表 8 的实验结果可以说明，本文提出的方法在虾类和香蕉评论数据的处理上都具有较好的性能。香蕉数据的实验结果在准确率上比虾类数据较低，但其召回率和 F 值都相对较高。由此也证明了该方法的通用性。

5 结论

本文提出了一种基于条件随机场模型的评价对象缺省项识别方法。首先通过 HowNet 情感词典加入词性信息提高观点句识别的准确性，并将识别评价对象缺省项位置的问题转换为序列标注问题，判断观点句中每个词之前是否存在缺省项，并结合了词法特征和句法特征，利用条件随机场模型进行标注。最后经过实验对方法性能进行测试，准确率达到了 86.03%，验证了本文方法的有效性与准确性。

在以后的研究中考虑扩展出更多的特征对性能进行改进。另外，由于商品评论的简短、口语化、不规范、缺省现象严重等特点，对商品评论对象恢复工作增加了困难。在以后的研究工作中，利用识别评价对象缺省项的位置帮助进行评价对象缺省恢复，以此来提高电商评论情感分析的性能的研究将成为重点。

参考文献

[1] 秦凯伟, 孔芳, 李培峰, 等. 基于规则的中文零指代项识别研究[J]. 计算机科学, 2012, 39(10): 278-281.

[2] Yeh C L, Chen Y C. Zero Anaphora Resolution in Chinese with Shallow Parsing[J]. Journal of Chinese Language and Computing, 2007, 17(1): 41-56.

[3] 杨国庆, 孔芳, 朱巧明, 等. 基于规则的中文缺省识别研究[J]. 计算机科学, 2011, 38(12): 255-257.

[4] Qin K, Kong F, Li P, et al. Chinese zero anaphor detection: rule-based approach[M].Knowledge Engineering and Management. Springer Berlin Heidelberg, 2011: 403-407.

[5] Zhao S, Ng H T. Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach[C]//EMNLP-CoNLL. 2007, 2007: 541-550.

[6] Kong F, Zhou G. A tree kernel-based unified framework for Chinese zero anaphora resolution[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010: 882-891.

[7] Song Yang, Wang Houfeng. Chinese Zero Anaphora Resolution with Markov Logic[J]. Journal of Computer Research and Development, 2015, 52(9): 2114-2122.

[8] 秦凯伟, 孔芳, 李培峰, 等. 用于中文缺省识别研究的机器学习方法[J]. Computer Engineering, 2012, 38(22):130-132.

[9] 刘慧慧, 王素格, 赵策力. 观点句中评价对象/属性的缺省项识别方法研究[J]. 中文信息学报, 2014,

28(6): 175-182.

[10] Yang Y, Xue N. Chasing the ghost: recovering empty categories in the Chinese Treebank[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010: 1382-1390.

[11] Rao S, Ettinger A, Hal Daumé I I I, et al. Dialogue focus tracking for zero pronoun resolution[C]//Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). 2015: 494-502.

[12] Chen C, Ng V. Chinese Zero Pronoun Resolution: A Joint Unsupervised Discourse-Aware Model Rivaling State-of-the-Art Resolvers[C]//Meeting of the Association for Computational Linguistics, 2015.

[13] Xue N, Xia F, Huang S, et al. The bracketing guidelines for the Penn Chinese Treebank (3.0)[J]. 2000.

[14] Yeh C L, Chen Y J. An Empirical Study of Zero Anaphora Resolution in Chinese Based on Centering Model[C]//ROCLING. 2001.

[15] Lafferty J, Mccallum A, Pereira F, et al. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]//International Conference on Machine Learning, 2001.

唐文武（1992--），硕士研究生，主要研究方向为自然语言处理、情感计算。

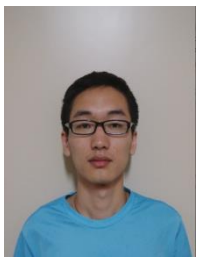
Email: tangww10101458@163.com

过弋（1975--），通讯作者，教授，博士，主要研究方向为自然语言处理、智能信息处理、本体工程。

Email: yguo1110@ecust.edu.cn

徐永斌（1990--），硕士研究生，主要研究方向为自然语言处理。

Email：xyb0723@sina.cn



方旭（1993--），硕士研究生，主要研究方向为自然语言处理。

Email：18817337228@163.com

作者联系方式：唐文武，上海市徐汇区梅陇路 130 号，200237，18801954785，tangww10101458@163.com; 过弋（通讯作者），上海市徐汇区梅陇路 130 号，200237，15821071625，yguo1110@ecust.edu.cn; 徐永斌，上海市徐汇区梅陇路 130 号，200237，xyb0723@sina.cn；方旭，上海市徐汇区梅陇路 130 号，200237，18817337228@163.com。