北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2015.046

中文文本中评价对象省略识别方法

朱珠　汪蓉　李寿山†　周国栋

苏州大学自然语言处理实验室，苏州　215006; † 通信作者, E-mail: shoushan.li@gmail.com

摘要　针对中文情感文本中频繁出现评价对象省略的情况，重点研究中文文本中评价对象省略识别现象。将评价对象省略识别建模为一个二元分类问题，利用机器学习算法进行自动学习。探讨当前句位置无关特征、当前句位置相关特征和上下文相关特征这 3 种不同类型的特征对评价对象省略识别的作用。3 个不同领域的实验结果表明，新提出的基于机器学习的评价对象省略识别方法能够获得较好的识别效果。

关键词　情感分析；评价对象抽取；评价对象省略；特征选择

中图分类号　TP391

Recognizing the Ellipsis of Opinion Target in Chinese Texts

ZHU Zhu, WANG Rong, LI Shoushan†, ZHOU Guodong

Natural Language Processing Laboratory, Soochow University, Suzhou 215006; † Corresponding author, E-mail: shoushan.li@gmail.com

Abstract　A novel method is proposed to recognize the ellipsis of opinion target in Chinese text by fully considering the common phenomenon in the ellipsis of opinion target in Chinese sentiment texts. The proposed approach treats the task of opinion target ellipsis as a binary classification problem, which applies the machine learning algorithm. Then three kinds of features, namely position-independent features of sentence, position-dependent features of sentence and contextual features are applied to the recognition task separately. The experimental results in three domains demonstrate that the machine learning-based method is effective for the task of the recognition of opinion target ellipsis.

Key words　sentiment analysis; opinion target extraction; ellipsis of opinion target; feature selection

随着现代网络的迅猛发展，互联网上涌现越来越多的对于人物、事件、产品等进行评论的主观性文本信息。因此，如何自动化地、智能化地获取和处理这些主观信息文本就显得极为重要，情感分析便是在这样的背景下产生并迅速发展的[1−3]。

情感信息抽取是情感分析的一个核心任务，要求从用户的论述中提取具有重要价值的信息(例如观点持有者、评价对象、情感表述等)，是一种细粒度的情感分析任务。评价对象抽取作为其中研究最为广泛的一项任务，有助于为上层的情感分析任务提供服务。评价对象指某段评论中情感表达所面向的对象，具体表现为评论文本中评价词语所修饰的对象。例如在产品评论中，评价对象常表现为产品本身(如“我很喜欢这款手机。”)或某一具体属性(如“polo 的外观很时尚。”)。

实际生活中，在不影响意思表达的前提下，人们通常会为了方便或者使语言简洁明快而省略部分信息。对于省略的信息，人们可以通过分析上下文或文本所处的环境而获知，但是机器却很难理解。省略现象在各种语言中都存在，由于中文本身的特点，中文文本中省略现象尤为突出。对中文省略的研究能够为其他自然语言处理任务提供帮助，尤其是在信息抽取任务中，文本缺失的信息可能正是用户所关心的。

国家自然科学基金(61375073, 61273320, 61331011)和 863 计划(2012AA011102)资助
收稿日期: 2014−07−26; 修回日期: 2014−10−11; 网络出版时间: 2014-11-28 15:22

在中文情感文本中，评价对象作为情感信息的一项重要元素，经常作为被省略的对象，这无疑会在一定程度上影响情感信息抽取系统的性能。如例 1 中虽然表达了正面的情感倾向，但是却省略了评价对象。又如例 2 中，分句“但是不实用。”省略了“不实用”的评价对象“这台笔记本”。目前有关中文文本中的评价对象省略现象并没有针对性的研究，而这种省略了评价对象的文本用以往的抽取方法无法达到判断评价对象的目的。

例 1 我很喜欢，很好看。

例 2 这台笔记本虽然好看，但是不实用。

因此，本文提出中文文本中评价对象省略识别方法。该任务的开展可以为情感分析任务或其他自然语言处理任务提供服务。首先，评价对象省略现象研究能够帮助提高现有评价对象抽取系统的性能，为省略评价对象的恢复提供服务。其次，随着新型社交网络的兴起，中文情感文本更加趋于口语化及表达的不完整化，评价对象的省略识别能够更加有效地帮助这些不规范文本的自动分析。最后，评价对象的省略作为中文文本省略现象的一个特殊情况，其研究可以为中文省略的相关研究提供新的思路。

本文将评价对象省略现象建模为一个二元分类问题，提出当前句位置无关特征、当前句位置相关特征和上下文相关特征这 3 种不同类型的特征，并将其应用到机器学习算法中。然后，使用贪婪式的特征选择算法，选取每个领域的特征集合。实验结果表明，本文提出的中文文本中评价对象省略识别方法能够获得较好的识别效果。

1 相关研究

1.1 评价对象抽取

评价对象抽取任务具有丰富的理论价值和应用价值，近年来许多学者对其进行研究，先后探索了很多不同的抽取方法。这些方法可以归纳为两个方向，即基于非监督学习的评价对象抽取方法和基于监督学习的评价对象抽取方法。

Hu 等[4]最早提出评价对象抽取问题，并使用词频、同情感词的距离等特征构建识别评价对象的启发式规则，实现了一种基于非监督学习的评价对象抽取方法。之后一些研究者陆续对非监督学习方法做了改进，如 Li 等[5]通过抽取〈情感词，评价对象〉二元组来捕获情感词和评价对象之间的关系，充分利用上下文信息来提高评价对象的抽取性能。

基于监督学习的评价对象抽取方法起步较晚，但是由于其更好的独立性和更优异的效果而逐渐占据主流。Zhuang 等[6]首先提出一种基于监督学习的评价对象抽取方法，从标注语料中生成评价对象表和情感词表，利用依存路径信息来识别语句中的<情感表达,评价对象>序偶，实验表明该方法明显优于其他基于非监督学习的方法。Jakob 等[7]将评价对象抽取建模为序列标注问题，构建基于条件随机场模型的抽取系统，在同一领域中获得比 Zhuang 等[6]的方法更优的抽取效果，同时也验证了该模型同样适用于评价对象抽取的领域适应任务。Li 等[8]将评价对象抽取问题转化成浅层语义分析任务，利用句法树进行监督学习，在充分利用句法信息的基础上获得更优的性能。

现有的评价对象抽取方法都是针对文本中明确表述的评价对象。考虑到中文复杂的语法和特殊的表达，本文将考虑中文文本中评价对象省略识别，以推进评价对象抽取的进一步研究。

1.2 中文省略

中文和英文在语法上有较大的区别，英文的省略严格按照语法产生。因此，英文发生省略时能够很好地利用语法信息识别出来。然而，由于中文语法情况更复杂，省略方式具有灵活性和不确定性，使其省略难以识别。目前中文省略的相关研究还处于起步阶段，关于情感文本中的评价对象省略现象也没有相关的研究。但是，有关零指代的研究较为成熟，并且与中文省略密切相关，其主要方法可以概括为基于规则的方法和基于机器学习的方法。

基于规则的方法主要利用语法信息判断语句中某个位置是否存在省略。Yeh 等[9-10]提出使用规则三元组的方法进行中文零指代消解的方案，为后续规则方法的研究提供了基础。杨国庆等[11]在 Yeh 等[10]方法的基础上将三元组改为五元组，实验结果显示他们的方法能够获得较好的性能。Nielsen[12]认为动词的相关信息能够帮助识别省略信息，并使用规则方法，通过分析 Parser 树识别 VPE (verb phrase ellipsis)的省略。

基于机器学习的方法主要是使用机器学习算法通过对标注样本的学习来构建分类器。Zhao 等[13]首先将机器学习方法应用到中文零指代消解任务中。Cui 等[14]发现以从句为单位进行零指代消解处理时所获得的性能要高于整句。黄李伟等[15]提出

基于树核函数的零指代识别方法, 实验表明该方法获得的识别性能相比基于规则的方法有显著的提高。

本文使用基于机器学习的方法, 结合贪婪式的特征选择算法, 选择合适的特征构建中文文本的评价对象省略识别系统。

2 中文文本中评价对象省略识别方法

2.1 概述

本文将评价对象省略识别建模为一个二元分类问题, 提出一种基于机器学习的评价对象省略识别方法。

首先, 考虑情感文本中出现了情感词的情况。将每一个情感词所在的分句作为一个样本, 若此情感词所评价的对象在整句中出现(评价对象省略现象未发生), 则分类为 0; 若此情感词所评价的对象在整句中未出现(评价对象省略现象发生), 则分类为 1。然后, 通过综合分析语料中的评价对象省略现象, 分别考察 3 个类别的特征(即当前句位置无关特征、当前句位置相关特征和上下文相关特征对评价对象省略识别)的作用。最后, 通过特征选择获取能够有效提高识别性能的特征组合。本文提出的评价对象省略识别系统的框架结构如图 1 所示。

2.2 特征概述

机器学习算法的学习性能很大程度上依赖于特征的选取。因此, 选择有效合理的特征至关重要。

在评价对象抽取任务中, 词形和词性特征都是常用的基本特征。一方面是因为在某个特定的领域中, 词形特征往往会直接提供当前词是否是评价对象的信息。例如在宾馆领域中, “宾馆”或“如家”等词是评价对象的概率较高。另一方面, 一个词的词性包含丰富的信息量(如大部分评价对象为名词或代词, 大部分情感词为形容词, 等等)。因此, 我们考虑充分利用词形和词性等较易获得的特征。

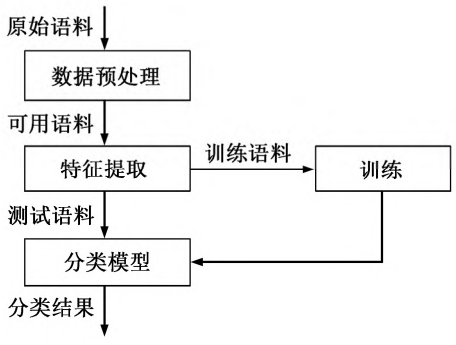

图 1 评价对象省略识别系统框架

Fig.1 Framework of recognizing the ellipsis of opinion target in Chinese text

本文提出并探索一些基于词形和词性的特征对中文评价对象省略识别的作用。这些特征大致可以分为 3 类: 当前句位置无关特征、当前句位置相关特征和上下文相关特征。

2.2.1 当前句位置相关特征

在选择特征的过程中, 考虑到当前句的位置能够提供一定的信息, 这些信息也许可以为省略现象的识别提供帮助。如评论文本“这本书非常好看。”中, 句首词“这本书”作为“好看”的评价对象的可能性较高, 进而可以判断情感词“好看”的评价对象没有省略。

当前句位置相关特征主要考虑利用句首词及句末词, 如表 1 所示。

2.2.2 当前句位置无关特征

主要考虑利用当前句的词形及词性特征。如评论文本“太贵了, 而且不好看。”中, 并没有出现名词或代词等可能是评价对象的词语, 进而可以初步判断情感词“贵”和“不好看”的评价对象被省略了, 如表 2 所示。

2.2.3 上下文相关特征

通常情况下, 上下文特征也能够提供一些信

表 1 当前句位置相关特征

Table 1 Position-dependent features of sentence

特征表示	特征说明
$f_i w$	当前句句首前 i 个词的词形特征
$f_i p$	当前句句首前 i 个词的词性特征
$f_i w\_p$	当前句句首前 i 个词的词形和词性的组合特征
$l_i w$	当前句句末最后 i 个词的词形特征
$l_i p$	当前句句末最后 i 个词的词性特征
$l_i w\_p$	当前句句末最后 i 个词的词形和词性的组合特征

说明: i∈[1, 当前句长度/2]。

表 2 当前句位置无关特征

Table 2 Position-independent features of sentence

特征表示	特征说明
Word	当前句的 bagword 特征
Pos	当前句所有词的词性特征
w_p	当前句所有词的词形和词性的组合特征

息。如情感词所在分句的上一个分句中存在的名词有可能是评价对象。例如在本文例 2 中, “但是不实用。”这一分句中“不实用”的评价对象“这台笔记本”便出现在其上一个分句“这台笔记本虽然好看, ”中。

因此本文考虑利用当前句的上下文信息, 选取当前句的前一个子句的相关信息为特征, 如表 3 所示。

3 实验设计与分析

3.1 语料设置

本文的实验语料来自于亚马逊网站(http://www.amazon.cn/)的产品评论, 其中包含 3 个领域, 分别为宾馆、笔记本和化妆品领域。我们对其中的评价对象和情感词部分进行了详细的标注, 其中与评价对象省略相关的统计信息如表 4 所示。

从表 4 可以发现, 省略评价对象在中文表达中较为常见, 在笔记本、宾馆和化妆品 3 个领域中, 分别有约 23%, 14%和 32%的句子发生评价对象省略现象。在进行实验之前, 需要对语料进行预处理, 使其适应系统的输入接口。实验前对语料的预处理主要有以下两步。

1) 获取词性信息。本文采用 Stanford Parser (http://nlp.stanford.edu/software/lex-parser.shtml#Citing)获取词的词性信息。

2) 抽取所需要的特征信息。将数据处理成分类模型所需的语料格式。下面以样本“真的很好用。”为例显示对其进行预处理之后的效果(为显示方便, 未列出实验所用的全部特征)。

真的 很 好 用 。 AD AD VA P PU 真的_AD 很_AD 好_VA 用_P 。_PU 真的_1 AD_2 。_3 PU_4 NULL

表 3 上下文相关特征

Table 3 Contextual features

特征表示	特征说明
PreW	当前句前一个子句的 bagword 特征
PreP	当前句前一个子句的所有词的词性特征
PreW_*P*	当前句前一个子句的所有词的词形和词性的组合特征

3.2 实验设置

本文采用最大熵分类模型作为分类算法, 在 3 个领域的语料上分别进行以下两个实验: 1) 针对 3.1 节提出的 3 类特征, 依次将单个特征加入系统进行实验; 2) 针对每个领域, 使用贪婪式的特征选择算法进行特征选择实验。

由于正负类样本存在明显的不平衡现象, 为了保证实验数据的准确性, 本文在采样过程中, 对负类样本进行随机欠采样, 最终结果使用 5 次随机采样后的平均结果。评价指标采用 F1 值(F1-Measure)。

3.3 实验结果与分析

3.3.1 加入单个特征后的识别结果

表 5 显示当前句位置无关特征在 3 个领域的识别结果。由表 5 可知, 此类别的 3 个特征在 3 个领域都取得很好的效果, 特别是 word 特征在化妆品领域的 F1 值达到 76.30%。这可能是由于评价对象与情感词在较多情况下都出现在同一子句中, 而词形及词性特征在一定程度上能够判断句中是否存在评价对象, 进而对判断句中情感词所评价的对象是否在子句中出现起到了很好的指示作用。

表 5 当前句位置无关特征的单个特征识别结果

Table 5 Performance of position-independent features of sentence separately

特征	F1/%		
	笔记本	宾馆	化妆品
word	75.0	72.4	76.3
pos	64.3	70.5	66.7
w_p	75.2	71.9	75.9

表 4 语料信息统计

Table 4 Statistics of data sets

领域	总篇章数	句子数	评价对象		情感词		含省略评价对象的句子数
			个数	文档平均个数	个数	文档平均个数	
笔记本	2000	4649	5167	2.85	6512	3.26	1082
宾馆	1000	4368	4864	4.86	5369	5.37	614
化妆品	2000	3291	2458	1.22	4359	2.18	1060

表 6 给出 i 分别取 1, 2, 3 时当前句位置相关特征在 3 个领域的识别结果。可以看出, 在仅使用单个特征的情况下, 句首第一个词和句末最后一个词的相关特征都有不错的效果。这说明当前句位置相关的信息对识别是否有评价对象省略现象是有效的。随着句首或者句末抽取特征所用词的数量增多, 并没有获得更好的效果。但是, 仔细观察可以发现, 剩余的位置相关特征中, 句首句末几个词的词性组合特征相较于其他特征, 其结果还是有一定程度的提高, 尤其是 $f_2\,p$ 特征(句首前两个词的词性特征)。由于很多时候句首几个名词或者代词就是评价对象, 故这样的结果也是比较合理的。

表 7 显示上下文相关特征在 3 个领域的识别结果, 3 个特征在笔记本领域的效果均低于随机结果, 在其他两个领域的效果略高于随机结果。由此, 可以初步判断单独使用上下文相关特征不能取得理想的效果。

3.3.2 特征组合后的识别结果

为了选取有效的特征组合, 更好地提高系统整体性能, 本文采用贪婪式的特征选择算法[16-17]。该算法选择剩余特征中效果最佳的特征加入系统。

通过贪婪式特征选择算法, 我们获取 3 个领域的最优特征集, 其特征选择结果如表 8~10 所示。

由表 8~10 可知, 经过贪婪式特征选择算法选择有效的特征组合后, 3 个领域的识别结果均达到 78%以上。特别是在宾馆领域, 识别性能达到 80%。该结果表明本文提出的方法能够有效的识别情感文本中是否发生了评价对象省略的情况。

在本文提出的 3 类特征中, 当前句位置无关特征在 3 个领域均能够入选特征组合, 说明当前句子本身的词形及词性特征能够对该任务产生很好的效

表 7 上下文相关特征的单个特征识别结果

Table 7 Performance of contextual features separately

特征	F1/%		
	笔记本	宾馆	化妆品
PreW	46.5	53.8	55.2
PreP	46.5	52.3	54.8
PreW_*P*	46.8	54.6	54.8

表 8 笔记本领域的特征选择结果

Table 8 Performance of feature selection in Notebook

加入特征	F1/%
w_p	71.9
word	75.8
f_2p	76.6
f_1w	77.1
pos	77.5
f_3p	77.7
l_1p	78.3

表 9 宾馆领域的特征选择结果

Table 9 Performance of feature selection in Hotel

加入特征	F1/%
word	72.4
w_p	75.4
pos	76.9
f_2p	77.7
f_1p	78.2
f_1w_p	78.7
f_1w	79.0
PreW_*P*	80.1

表 6 当前句位置相关特征的单个特征识别结果

Table 6 Performance of position-dependent features of sentence separately

特征	F1/%			特征	F1/%			特征	F1/%		
	笔记本	宾馆	化妆品		笔记本	宾馆	化妆品		笔记本	宾馆	化妆品
f_1w	70.4	66.9	69.2	f_2w	54.7	52.2	56.4	f_3w	38.6	36.7	39.5
f_1p	66.8	69.8	62.6	f_2p	73.0	70.3	69.1	f_3p	63.7	64.1	63.6
f_1w_p	69.2	67.5	67.9	f_2w_p	53.6	50.1	51.9	f_3w_p	40.3	35.5	38.6
l_1w	66.1	61.8	61.9	l_2w	57.6	54.7	54.5	l_3w	41.7	39.8	38.6
l_1p	61.7	53.0	55.2	l_2p	62.1	55.9	59.4	l_3p	58.8	64.4	61.6
l_1w_p	64.8	62.0	59.0	l_2w_p	55.8	52.9	53.6	l_3w_p	42.1	40.2	43.3

说明: $i=1, 2, 3$。

表 10　化妆品领域的特征选择结果

Table 10　Performance of feature selection in Beauty

加入特征	F1/%
word	76.3
w_p	76.4
$f_1 w$	76.6
$f_2 p$	77.2
pos	77.5
$f_1 p$	77.7
$l_3 p$	77.9
$l_1 w\_p$	78.3
PreW	78.7

果，并具有一定程度的领域适应性。在当前句位置有关特征中，除 f_1 *w* 特征外，主要的有用特征都集中在词性的相关特征上，如 f_1 *p* 和 f_2 *p* 等。上下文相关特征虽然在单个特征的实验结果上不能取得较好的效果，但在特征组合过程中却表现出一定效果，特别是在宾馆领域中，PreW_*P* 特征的加入使系统的整体性能提高 1.1%。经分析发现，这可能是由于宾馆文本较其他两个领域的文本篇幅更长，上下文能够提供的有用信息更多。

4　小结

本文提出中文文本中评价对象省略识别方法。该方法将评价对象省略现象建模为一个二元分类问题，使用机器学习方法对情感文本进行评价对象省略识别。在特征方面，提出并考察了当前句位置无关特征、当前句位置相关特征和上下文相关特征这 3 种不同类型的特征对识别系统的作用。实验结果表明，本文提出的基于机器学习方法的评价对象省略识别方法能够获得较好的识别效果，最终选择的特征集合在 3 个不同领域都能够达到 80%左右的识别性能。同时，我们发现在不同领域的最优特征集合中有较多特征发生重合，表明这些特征具有一定程度的领域适应性。

在下一步工作中，我们将尝试探索更多更有效的特征，以提高评价对象省略的识别性能，并进一步考虑将评价对象省略识别方法应用到评价对象抽取任务中，以期进一步提高评价对象的抽取性能。

参考文献

[1] Pang B, Lee L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2008, 2: 1−135

[2] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques // Proceedings of EMNLP-2002. Philadelphia, 2002: 79−86

[3] 赵妍妍, 秦兵, 刘挺. 文本情感分析. 软件学报, 2010, 21(8): 1834−1848

[4] Hu M, Liu B. Mining opinion features in customer reviews // Proceedings of AAAI-2004. San Jose, 2004: 755−760

[5] Li B, Zhou L, Feng S, et al. A unified graph model for sentence-based opinion retrieval // Proceedings of ACL-2010. Uppsala, 2010: 1367−1375

[6] Zhuang L, Jing F, Zhu X. Movie review mining and summarization // Proceedings of CIKM-2006. Virginia, 2006: 43−50

[7] Jakob N, Gurevych I. Extracting opinion targets in a single and cross-domain setting with conditional random fields // Proceedings of EMNLP-2010. Massachusetts, 2010: 1035−1045

[8] Li S, Wang R, Zhou G. Opinion target extraction using a shallow semantic parsing framework // Proceedings of AAAI-2012. Toronto, 2012: 1671−1677

[9] Yeh C, Chen Y. Zero anaphora resolution in chinese with shallow parsing. Journal of Chinese Language and Computing, 2007, 17(1): 41–56

[10] Yeh C, Mellish C. An empirical study on the generation of anaphora in Chinese. Computational Linguistics, 1997, 23(1): 171–190

[11] 杨国庆, 孔芳, 朱巧明. 基于规则的中文缺省识别研究. 计算机科学, 2012, 38(12): 255–257

[12] Nielsen L. Verb phrase ellipsis detection using automatically parsed text // Proceedings of COLING-2004. Geneva, 2004: 1093–1100

[13] Zhao S, Ng H. Identification and resolution of chinese zero pronouns: a machine learning approach // Proceedings of EMNLP-CoNLL-2007. Prague, 2007: 541–550

[14] Cui Y, Hu Q, Pan H, et al. Zero anaphora resolutionin Chinese discourse. Computational Linguistics and Intelligent Text Processing, 2006: 245–248

[15] 黄李伟, 孔芳, 朱巧明, 等. 基于树核函数的中文零指代项识别研究. 计算机科学, 2011, 38(1): 214–216

[16] Jiang Z, Ng H. Semantic role labeling of nom bank: a maximum entropy approach // Proceedings of EMNLP-2006. Sydney, 2006: 138–145

[17] Ding W, Chang B. Improving Chinese semantic role classification with hierarchical feature selection strategy // Proceedings of EMNLP-2008. Honolulu, 2008: 324–333