评价对象抽取研究综述

蒋盛益 [1,2] 郭林东 [1] 王连喜 [3] 符斯慧 [1]

摘 要 近年来, 细粒度情感分析因其在商业决策、舆情分析等领域的重要作用而受到学术界和工业界的广泛关注. 评价对象抽取作为情感分析的基本任务之一, 是进行细粒度情感分析的关键问题. 本文针对评价对象抽取问题的起源、当前主流研究方法和趋势进行了梳理, 首先详细阐述评价对象抽取问题的基本概念并对其进行形式化表示, 然后结合近年来的研究对评价对象抽取方法进行归纳和总结, 并重点分析基于频率、基于模板规则、基于图论、基于条件随机场和基于深度学习的评价对象抽取方法, 随后回顾评价对象抽取的评测情况和可用的语料资源, 最后分析评价对象抽取的若干难点问题, 同时对评价对象抽取研究进展和发展趋势进行总结和展望.

关键词 评价对象抽取, 细粒度情感分析, 评测, 资源建设

引用格式 蒋盛益, 郭林东, 王连喜, 符斯慧. 评价对象抽取研究综述. 自动化学报, 2017, **43**(X): X−X

DOI 10.16383/j.aas.2017.c170049

Survey on Opinion Target Extraction

JIANG Sheng-Yi[1,2] GUO Lin-Dong[1] WANG Lian-Xi[3] FU Si-Hui[1]

Abstract In recent years, fine-grained sentiment analysis has received extensive attention from academia and industry for its important role in business decision-making and public opinion analysis. Opinion target extraction, as one of the basic tasks of sentiment analysis, has been the crux of fine-grained sentiment analysis for years. In this paper, the origins, state of the art, and research directions are discussed. We first elaborate the basic concepts of opinion target extraction and formalize the question, and then based on recent literature, we conclude and summarize the approaches and techniques which we divided to five categories: frequency-based method, pattern-based method, graph-based method, CRF-based method and deep learning based method. We review several evaluation contests and collect the available corpus resources on opinion target extraction. At the end, the challenges arisen on the opinion target extraction are analyzed as well as the probable future are given.

Key words opinion target extraction, fine-grained sentiment analysis, evaluation, resource construction

Citation JIANG Sheng-Yi, GUO Lin-Dong, WANG Lian-Xi, FU Si-Hui. Survey on Opinion Target Extraction. *Acta Automatica Sinica*, 2017, **43**(X): X−X

评价对象抽取是情感分析 (亦称意见挖掘[1−2]) 中的重要问题, 例如在产品评论中, "这手机的屏幕挺好的, 就是电池太不耐用了", "屏幕" 和 "电池" 均为评价对象, 该评论对屏幕的情感倾向体现为正向, 而对电池的情感倾向体现为负向, 但是按照传统的情感分析方法, 如果将评论文本看作一个整体, 便会造成分析结果准确率下降甚至错误. 又如, 在新闻评论中, "可怜的孩子呀 一定要将罪犯绳之于法!", 评论者对孩子的情感为同情, 对罪犯的情感为憎恨, 评论文本对不同评价对象表达了不同的情感, 因此不能对整个句子进行情感分析, 需要首先抽取出句子中的评价对象, 然后再判别针对评价对象的情感倾向. 评价对象抽取的有效解决有利于挖掘出不同对象的情感, 有助于对产品或公共事件进行细粒度的情感分析.

本文立足于国内外现有研究成果, 对评价对象抽取进行综述. 第 1 节介绍评价对象的定义和抽取的形式化表示; 第 2 节按照研究者使用的主要技术

收稿日期 2017-01-19 录用日期 2017-04-17
Manuscript received January 19, 2017; accepted April 7, 2017
国家自然科学基金 (61572145), 教育部人文社会科学青年项目 (14YJC870021), 广东省科技计划项目 (2014A040401083, 2015A030401093), 广东省大学生科技培育专项资金项目 (110-GK161017) 资助
Supported by National Natural Science Foundation of China (61572145), Youth Project of Humanities and Social Sciences of the Ministry of Education (14YJC870021), Science and Technology Planning Project of Guangdong Province (2014A040401083, 2015A030401093), Guangdong College Students in Science and Technology Innovation and Cultivation of Special Funds(110-GK161017)
本文责任编委 赵铁军
Recommended by Associate Editor ZHAO Tie-Jun
1. 广东外语外贸大学信息科学与技术学院 广州 510006 2. 语言工程与计算广东省社会科学重点实验室 广州 510006 3. 广东外语外贸大学图书馆 广州 510420
1. School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou 510006 2. Social Science Key Laboratory of Language Engineering and Computing of Guangdong Province, Guangzhou 510006 3. Guangdong University of Foreign Studies Library, Guangzhou 510420

对评价对象抽取方法进行详细归纳和梳理; 第 3 节介绍相关评测和语料资源; 第 4 节剖析评价对象抽取的难点问题; 第 5 节对领域内的研究进展进行总结, 并对未来的发展趋势进行展望.

1 问题定义与形式化表示

1.1 问题定义

在国际著名评测 SemEval 中, 将评价对象定义为能够用来指代给定文本中评价实体特征的表达方式. 按照该定义的描述, 评价实体可以是产品, 也可以是抽象概念. 当评价实体为产品时, 例如手机、笔记本等, 评价对象可以是产品特征、产品功能或产品零件; 当评价实体为抽象概念时, 例如新闻事件等, 评价对象往往是人、组织机构或利益集团. 从语言学角度看, 评价对象通常是名词或名词短语. 以下通过具体示例进行分析:

例 1 The noise level was unbearable, conversation impossible. (from="4" to="15")

例 2 Highly recommended is the Spicy Fried Clam Rolls and Spider Rolls. (from="26" to="48"; from="53" to="65")

例 3 Great for a romantic evening, but overpriced.

例 1 中, 评价对象为 "noise level", 是名词短语, 由两个单词组成, 其中 from 和 to 指评价对象在句子中的位置 (从 0 开始); 例 2 中, "Spicy Fried Clam Rolls" 和 "Spider Rolls" 均为评价对象, 它们是并列关系, 共享同一情感评价词; 例 3 中评价对象没有显式出现在文本中, 但根据上下文语义可知评价对象是 "price", 称为隐式评价对象. 因为隐式评价对象抽取难度较大, 所以其研究相对较少: 对同一评论文本, 不同的人对其中隐含评价对象会有不同的看法, 往往很难通过实验评测界定; 此外, 有些隐式评价对象通过隐喻或反讽进行表达, 需要事先了解背景知识才能做出判断.

评价对象抽取问题容易与命名实体识别问题混淆, 两者既有相同点, 也有区别. 两者都可以看作序列标注问题, 但后者重点在于识别出待处理文本中出现的人名、地名和机构名等专有名词, 评价对象只出现在主观性句子中而不出现在陈述句中.

1.2 形式化表示

针对任务和粒度的不同, 评价对象抽取可以划分为语料级别抽取和语句级别抽取. 语料级别评价对象抽取旨在获取一个评价对象集合, 例如关于一个产品的所有评价对象 (整体、属性、部件、部件属性等), 关于一个新闻话题或事件所涉及的所有人物、机构等利益团体. 语句级别评价对象抽取则侧重于识别出一个句子单位中提及的评价对象. 它们的区别如图 1 所示. 前者适用于产品领域的观点摘要等应用级任务; 后者适用于对不同评价对象的观点进行精确统计分析, 如在新闻事件中,

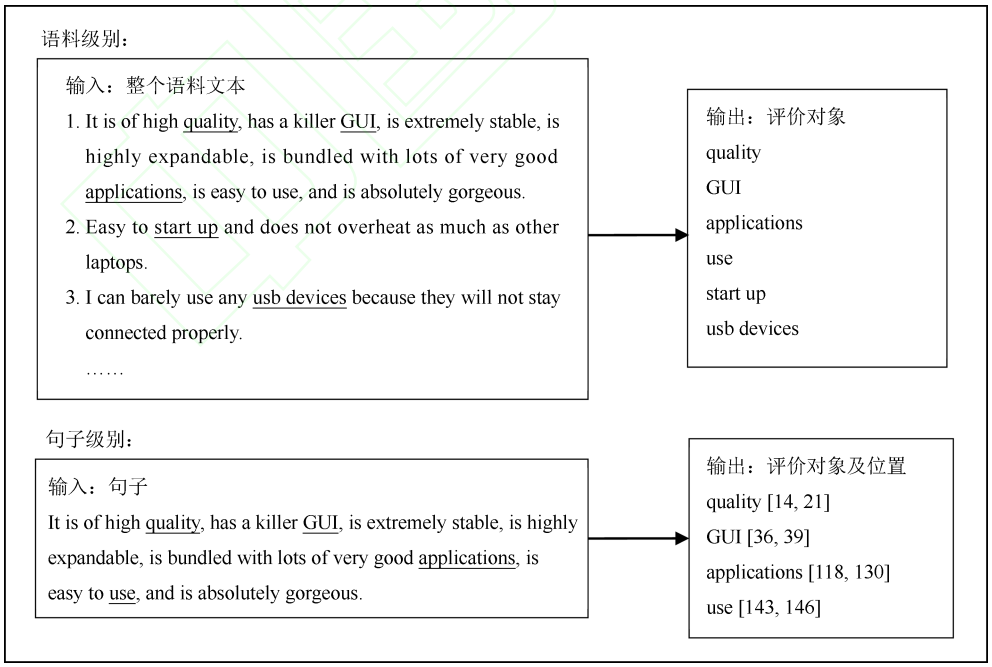

图 1 语料级别和句子级别任务的区别

Fig. 1 Difference between corpus level task and sentence level task

很可能需要统计对于每个评价对象持正负向观点的数目. 这里重点介绍两者的形式化表示, 对于语料级别评价对象抽取, 形式化表示为: 给定语料 $C = \{r_1, r_2, r_3, \cdots, r_m\}$, 其中 m 为评论数量, r_m 为第 m 条评论, 任务的目标是根据 C 得到评价对象词语集合 $T = \{t_1, t_2, t_3, \cdots, t_n\}$, 其中 n 为得到的评价对象数量, t_n 表示第 n 个评价对象, 评价对象可以是一个词或包含一个词以上的短语; 语句级别评价对象抽取, 则形式化表示为: 给定语句 $S = \{w_1, w_2, w_3, \cdots, w_m\}$, 其中 m 为词语数量, w_m 表示语句中第 m 个词, 任务的目标是由 S 得到评价对象元组集合 $T = \{< t_1, p_1 >, < t_2, p_2 >, < t_3, p_3 > \cdots < t_n, p_n >\}$, 其中 n 为评价对象数量, t 为评价对象词语, p 为评价对象在语句中的位置信息, $< t_n, p_n >$ 为第 n 个评价对象元组.

语料级别抽取一般可以使用无监督学习方法进行识别, 而无监督方法通常具有领域无关、语言无关的优势, 因此可以很方便地迁移到其它领域或语言上. 无监督评价对象抽取方法在某些场合下非常有用, 例如给定一个没有经过标注的话题事件新闻和对应的网民评论, 通过此方法可以得到该事件主要涉及的评价对象. 其缺点是难以确定最终评价对象数目的阈值和容易混入噪音. 语句级别抽取通常把问题看作是序列标注问题, 优点在于能精确抽取出评价对象及其在句子中的位置, 这种方法依赖于评价对象在句子中的上下文信息, 对于缺少语境信息的短评, 这种方法难以取得好的效果.

2　评价对象抽取研究方法

自评价对象抽取问题提出 10 多年以来, 文献中已经出现了许多方法和模型, 从最开始的基于频率的方法到近年来的基于深度学习方法, 评价对象抽取的性能获得大幅提升. 本文按照研究者使用的主要技术将他们分为基于频率、基于模板规则、基于图论、基于条件随机场和基于深度学习的方法, 总结归纳如图 2 所示. 下面将对其进行详细阐述.

2.1　基于频率的方法

由于评价对象往往是评论文本中的名词和名词短语, 而人们在对产品或新闻事件进行评论时, 描述评价对象的用语较为集中且在语料中频繁出现, 基于此假设提出了基于频率的抽取方法: 通过频率统计方法抽取评论文本中的评价对象. Hu 等最早使用关联规则方法识别评价对象[3], 该方法的流程图如图 3 所示, 对于频繁评价对象, 首先将产品评论文本中的名词或名词短语看作是候选评价对象, 然后利用关联规则算法抽取频繁评价对象, 之后, 利用剪枝方法进行评价对象的筛选; 对于非频繁的评价对象, 选取离评价词位置最近的名词或名词短语作为评价对象. 这种方法简单易行, F 值达到 80%. 但对于频繁的评价对象, 容易引入较多噪音 (非评价对象), 如人们生活中的常用语, 造成准确率不高; 对于非频繁的评价对象抽取, 容易出现评价词缺失和远距离评价对象的情况, 造成召回率不高.

针对以上问题, 学者们设法寻找评价对象的词频、评价文本的语法语义等特点, 通过总结相关规律以提高准确率和召回率. Popescu 等[4] 利用点互信息 (Point-wise mutual information, PMI) 技术来评估每个候选评价对象, 从而剔除那些可能的非评价对象的词语, 在相同的数据集上, 比文献[3] 的方法在准确率上提高了 6%. 因为该算法通过计算候选评价对象和产品类别指示词 (Class-specific discriminator) 之间的 PMI 值, 所以需要事先知道产品类别信息. 同时该方法需要借助搜索引擎, 所以会增加耗费额外的时间开销. Blair-Goldensohn 等[5] 结合句法模板、相对词频和情感词等进行评价对象的抽取和筛选. 例如在句法模板方面, 如果形容词后面跟着名词或名词短语, 则该名词或名词短语很可能是评价对象.

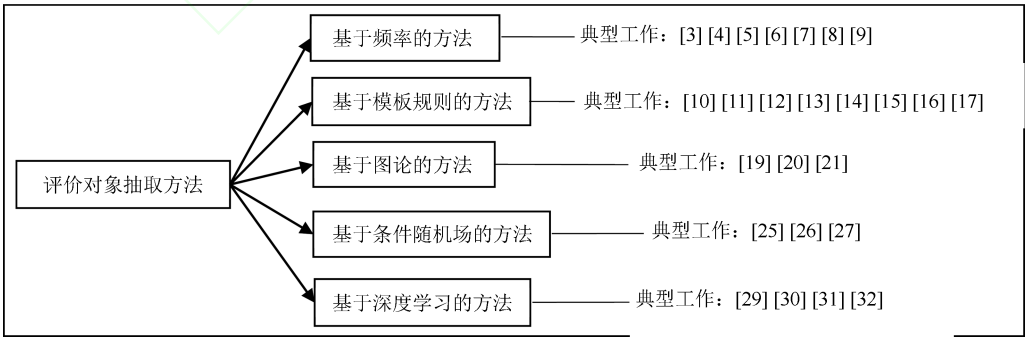

图 2　评价对象抽取研究方法概述

Fig. 2　Summary of opinion target extraction methods

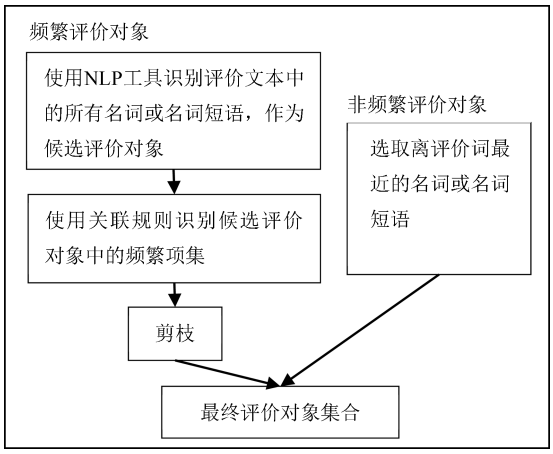

图 3 基于关联规则的抽取方法步骤

Fig. 3 Procedure of the extraction method based on association rule

Scaffidi 等[6] 假设评价对象在产品评论中出现的频率要高于一般语料, 提出了一种语言模型来解决产品评论文本评价对象抽取问题. 该方法基于统计模型, 对于出现次数较多的评价对象抽取效果会比较好, 但对于出现次数不多的评价对象效果不稳定, 并且该实验仅度量了准确率, 对于召回率情况并无描述.

当然, 还有一些其他的基于频率的方法, 如 Yi 等[7] 提出了基于混合语言模型和似然比检验 (Likelihood ratio test, LRT) 的方法. Xu 等[8] 将中文文本句子通过 Skip-Bigram 来表示, 避免了中文分词在产品评论文本中准确率较低的问题, 然后通过关联规则和筛选剪枝得到最终结果. Li 等[9] 在文献[3] 的基础上, 通过去除通用词、领域常用词, 采用基于词序的剪枝和基于语义相似度 (PMI-IR) 的剪枝改进性能. 基于频率的评价对象抽取方法总结归纳如表 1 所示.

2.2 基于模板规则的方法

基于模板规则的评价对象抽取方法通过观察评价对象特有的位置信息或评价对象与评价词之间的句法或语法关系, 然后构建词形模板、词性模板、依存关系模板或语义角色模板等进行评价对象抽取. 该类方法的优点在于抽取的准确率较高, 缺点是需要事先准备好评价词集合. 此外, 模板规则的制定、规则匹配的先后顺序和冲突问题也是基于模板规则方法的难点.

Zhuang 等[10] 在训练语料中使用 Stanford Parser 进行依存句法分析, 根据评价对象和评价词抽取出频率较高的依存关系模板, 然后使用预先定义好的评价对象和评价词列表, 结合依存关系模板对电影评论文本进行 “特征 - 观点” 对抽取. Jakob 等[11] 在文献[10] 方法的基础上, 改进了现有的指代消解技术来提高评价对象抽取的准确率. 宋晓雷等[12] 对汽车评论语料进行挖掘, 提出一种针对特定领域的无须借助外部资源的评价对象抽取方法, 该方法使用词性和词形模板、模糊匹配和剪枝法得到候选评价对象, 然后使用双向 Bootstrapping 方法筛选出评价对象, 最后使用 K 均值算法

表 1 基于频率的抽取方法比较

Table 1 Comparison of frequency-based extraction methods

文献	方法	级别	数据集	语言	实验结果 [1]
Hu 等[3] 2004	关联规则挖掘频繁项集 + 剪枝 + 根据评价词获取非频繁评价对象	语料	Digital Camera 1	English	0.747 0.822 0.783
			Digital Camera 2	English	0.710 0.792 0.749
			Cellular Phone	English	0.718 0.761 0.739
			MP3 Player	English	0.692 0.818 0.750
			DVD Player	English	0.743 0.797 0.770
Popescu 等[4] 2005	PMI Assessment	语料	Digital Camera 1	English	0.89 0.80 0.84
			Digital Camera 2	English	0.87 0.74 0.80
			Cellular Phone	English	0.89 0.74 0.81
			MP3 Player	English	0.86 0.80 0.83
			DVD Player	English	0.90 0.78 0.84
Xu 等[8] 2012	Skip-Bigram + 关联规则+ 剪枝	语料	Mobile Phone	Chinese	0.4081 0.9529 0.5715
			Digital Camera	Chinese	0.3828 0.7153 0.4987
Li 等[9] 2015	Hu 等[3] 的基础上+ 基于词序的过滤+ 基于PMI-IR 的过滤	语料	Mobile Phone 1	Chinese	0.732 0.667 0.698
			Mobile Phone 2	Chinese	0.791 0.850 0.819
			Digital Camera 1	Chinese	0.721 0.756 0.738
			Digital Camera 2	Chinese	0.719 0.639 0.676

[1] 实验结果分别表示精确率 (Precision)、召回率 (Recall) 和 F1 值.

区分产品名称和产品属性. 该方法简单易行, 特别在新领域语料不充分的情况下处理海量在线产品评论具有重要研究意义.

Qiu 等[13] 提出了一种叫做 Double Propagation (DP) 的非监督学习评价对象和评价词的抽取方法, 流程图如图 4 所示. 此方法根据预先定义好的依存关系模板来刻画评价词和评价对象之间的关系, 以此根据评价词来抽取评价对象, 根据评价对象抽取评价词, 然后根据评价词和评价对象来抽取新的评价词和评价对象, 依此循环, 直到没有新的评价词和评价对象. 最后, 利用剪枝方法对评价对象进行筛选提纯. 此方法无需外部资源, 仅需要少量种子词, 避免了监督学习的人工标注过程, 受到学者的广泛重视.

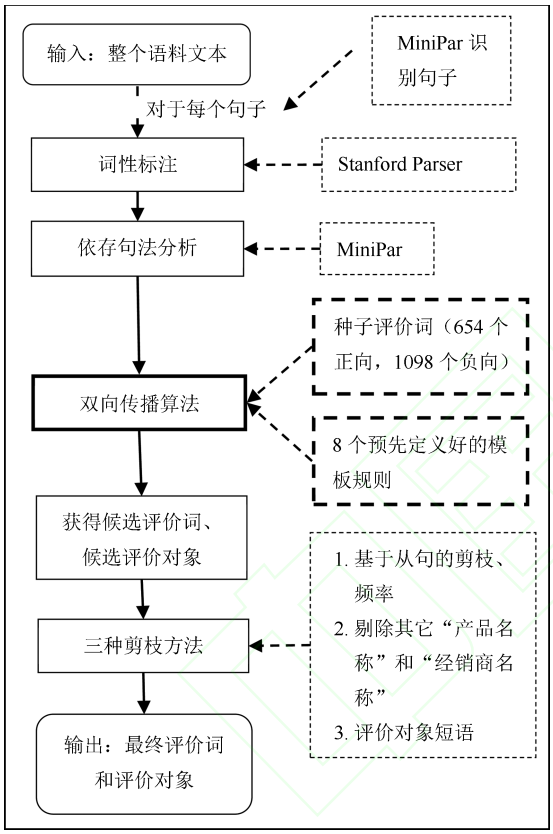

图 4　DP 算法进行评价对象抽取步骤

Fig. 4　Procedure of the DP algorithm

DP 算法在语料较大的情况下会引入较多无法剔除的噪音, 导致准确率降低, 而在语料较小的情况下, 召回率亦会有所下降. 针对此问题, Zhang 等[14] 首先增加 “部分 – 整体” 模板和 “No” 模板来提高召回率; 然后使用 HITs 算法进行评价对象的排序, 并将上一步获取的候选评价对象作为 Authority 节点, 特征指示词作为 Hub 节点, 其中特征指示词包括评价词、评价实体和 “No”, 以此构建二部图, 同时计算每个候选评价对象的权威值, 最后结合候选评价对象的频率得到最终得分, 得分超过一定阈值的即为最终的评价对象. 实验结果显示, 在语料较小时 (1 000 个句子), 在准确率稍微降低的情况下, 召回率得到较大提升; 在语料较大时 (3 000 个句子), 除了在手机领域准确率有所下降之外, 在其它领域中准确率和召回率都有所提升.

模板规则的构建通常受限于相似领域, 存在领域迁移困难的问题, 因此有学者研究规则的自动构建[15−16], 以解决领域适应问题. 赵妍妍等[15] 提出了句法路径库的自动构建方法, 并用于评价对象的识别, 该方法分为两个步骤, 一是从大量未标注语料中构建句法路径库, 二是使用基于编辑距离的句法路径匹配算法识别情感句中的情感评价单元. 实验结果表明, 句法路径频率为 70, 编辑距离为 1 时能获得最好的结果. 该方法的有效性表明评价词和评价对象在句法结构树中存在一定的关系, 而且这种关系在评论文本中普遍存在.

江腾蛟等[17] 研究了金融领域上市公司研究文本的 “评价对象 - 情感词” 对抽取. 首先构造了 3 947 个金融领域情感词, 然后对大量金融评论文本进行句法结构、依存关系、浅层语义分析后, 归纳了 8 种规则以实现根据金融领域情感词抽取其所修饰的评价对象. 运用特殊情感词评价搭配表、上下文语义关联评价搭配表和频繁评价搭配表以实现隐式评价对象的抽取. 实验结果显示, 准确率达到了 80 % 以上, 召回率为 97.3 %. 该方法在行文较为规范、语法单一的金融文本中能取得较好的效果, 而对于股吧评论等行文随意、句法语法复杂多变的文本则效果不佳. 基于模板规则的评价对象抽取方法总结归纳如表 2 所示.

2.3　基于图论的方法

基于图论的评价对象抽取方法考虑了评价词和评价对象之间的关系, 其基本流程如下: 首先把名词或名词短语加入到候选评价对象集, 把形容词加入到候选评价词集, 然后采用依存句法分析器或词对齐模型等技术捕获候选评价对象和候选评价词之间的关系, 以候选评价对象和候选评价词作为顶点, 候选评价词与候选评价对象之间的关系作为连边, 构建一个异构图. 最后, 在图上使用协同排序 (Co-ranking) 算法计算得到候选项 (候选评价对象和候选评价词) 的置信度, 置信度较高的项往往就是正确的评价对象或评价词. 基于图论的评价对象抽取方法总结归纳如表 3.

Liu 等[19] 最早使用基于图论和词对齐模型进行评价对象抽取. 首先通过词对齐模型挖掘候选评价对象和候选评价词之间的关系, 然后使用基于图论

表 2 基于模板规则的抽取方法比较

Table 2 Comparison of pattern-based extraction methods

文献	方法	级别	数据集	语言	实验结果 [1]
Zhuang 等[10] 2006	Dependency Template, Supervised	语料	IMDb Movie Reviews	English	0.483 0.585 0.529 [2]
刘鸿宇等[18] 2010	句法路径 + 词频过滤+ PMI 过滤+ 名词剪枝	句子	Car	Chinese	0.3126 0.4127 0.3557 (S) 0.4513 0.5958 0.5136 (L)
			Camera	Chinese	0.3219 0.4006 0.3569 (S) 0.4556 0.567 0.5052 (L)
			Phone	Chinese	0.3472 0.381 0.3633 (S) 0.4976 0.546 0.5206 (L)
			NoteBook	Chinese	0.3379 0.4566 0.3883 (S) 0.4693 0.6342 0.5394 (L)
Qiu 等[13] 2011	Double Propagation	语料	D1	English	0.87 0.81 0.84
			D2	English	0.90 0.81 0.85
			D3	English	0.90 0.86 0.88
			D4	English	0.81 0.84 0.82
			D5	English	0.92 0.86 0.89

[1] 实验结果中 S 表示 Strict 评测方式, L 表示 Lenient 评测方式.

[2] Based on feature-opinion pairs.

表 3 基于图论的抽取方法比较

Table 3 Comparison of graph-based extraction methods

文献	方法	级别	数据集	语言	实验结果
Liu 等[19] 2012	词对齐 + 随机游走	语料	Camera	Chinese	0.75 0.81 0.78
			Car	Chinese	0.71 0.71 0.71
			Laptop	Chinese	0.61 0.85 0.71
			Phone	Chinese	0.83 0.74 0.78
			Hotel	English	0.71 0.80 0.75
			MP3	English	0.70 0.82 0.76
			Restaurant	Chinese	0.80 0.84 0.82
			D1	English	0.84 0.85 0.84
			D2	English	0.87 0.85 0.86
			D3	English	0.88 0.89 0.88
			D4	English	0.81 0.85 0.83
			D5	English	0.89 0.87 0.88
Zhou 等[21] 2013	标签传播	句子	Tencent Weibo	Chinese	0.43 0.39 0.41 (Strict) 0.61 0.55 0.58 (Soft)
Xu 等[20] 2013	依存句法分析 + 随机游走 + 直推式支持向量机	语料	D1	English	0.86 0.82 0.84
			D2	English	0.88 0.83 0.85
			D3	English	0.89 0.86 0.87
			D4	English	0.83 0.86 0.84
			D5	English	0.89 0.85 0.87

的方法对候选评价对象进行评分, 选取评分较高的作为最终评价对象.

给定一个包含 n 个词的句子 $S = \{w_1, w_2, w_3, \cdots, w_n\}$, 词对齐集合 $A = \{(i, a_i) \mid i \in [1, n]\}$ 可通过最大化句子之间词对齐的概率获

得:

$$\hat{A} = \arg\max_{A} P(A \mid S)$$

其中 (i, a_i) 表示在第 i 个位置的名词或名词短语与在第 a_i 个位置的形容词的词对齐信息. 然后, 分别使用了三个 IBM 词对齐模型刻画他们之间的关系:

$$P_{IBM-1}(A|S) \propto \prod_{j=1}^{n} t(w_j \mid w_{a_j})$$

$$P_{IBM-2}(A|S) \propto \prod_{j=1}^{n} t(w_j \mid w_{a_j}) d(j \mid a_j, n)$$

$$P_{IMB-3}(A|S) \propto \prod_{i=1}^{n} n(\phi_i \mid w_i) \prod_{j=1}^{n} t(w_j \mid w_{a_j}) d(j \mid a_j, n)$$

其中, $t(w_j \mid w_{a_j})$ 刻画了名词或名词短语 w_j 与形容词 w_{a_j} 在语料中的共现关系; $d(j \mid a_j, n)$ 刻画了词之间的位置信息; $n(\phi_i \mid w_i)$ 刻画了词的搭配能力, ϕ_i 表示在句子中与词 w_i 对齐的词的数量. 三个模型是递进关系, 后一个模型是前一个模型的强化版本. 构造二部图时, 连边的权重刻画了名词或名词短语 (评价对象) 与形容词 (评价词) 之间的相关性, 顶点的权重则刻画了候选评价对象的置信度. 其中名词或名词短语与形容词之间的关系度量由 $Association(w_N, w_A)$ 给出, 顶点初始权重由 $Importance(c)$ 给出:

$$Association(w_N, w_A) = \frac{1}{\frac{t}{p(w_N, w_A)} + \frac{1-t}{p(w_A, w_N)}}$$

其中 t 为调和因子, $p(w_N, w_A) = \frac{Count(w_N, w_A)}{Count(w_A)}$, $p(w_A, w_N) = \frac{Count(w_A, w_N)}{Count(w_N)}$.

$$Importance(c) = \frac{tf - idf(c)}{\sum_c tf - idf(c)}$$

其中 c 为候选项, $tf - idf(c)$ 即为著名的 TF-IDF 公式, 用于刻画词的重要性.

二部图构建完毕后, 在二部图上使用基于随机游走算法, 收敛后可获得每个顶点的最终权重, 然后选择权重超过一定阈值的候选评价对象作为最终的评价对象. 作者分别在 COAE2008 Dataset 2, Large, CRD[3] 三个数据集上进行实验, 其中 Large 为作者搜集整理的数据集. 实验结果显示, 作者提出的模型相比文献[3]、[13] 和[14] 在准确率、召回率和 F 值上均有所提升.

Xu 等[20] 提出了一个二阶段框架来实现评价词和评价对象的抽取, 在第一阶段, 使用 MiniPar 和 Stanford Parser 对句子文本进行词性标注和依存句法分析, 将所有的形容词标记为候选评价词, 将所有的名词或名词短语标记为候选评价对象, 记录候选评价词和候选评价对象之间的最短依存路径, 之后, 利用上述产生的候选评价词、候选评价对象以及依存路径模板作为顶点, 构建三部图, 顶点之间的连边权值由函数 $w(v_a, v_b) = freq(v_a, v_b)/freq(v_b)$ 给出, 其中 v_a, v_b 为顶点, $freq(v_a, v_b)$ 表示共现频率, 然后在图上使用带重启的随机游走算法得到评价词和评价对象的顺序表, 在顺序表上排名靠前的很可能是正确的评价词和评价对象; 在第二阶段, 由于上一阶段挖掘出来的候选项会混有较多噪音, 比如某些出现频率很高但不是评价对象的词语, 或者某些出现频率不高, 但却是评价对象的词语, 即长尾评价对象, 作者在此阶段使用了直推式支持向量机算法解决这个问题.

除了利用评价词和评价对象之间的关系外, Zhou 等[21] 提出了一种非监督标签传播算法用来抽取微博文本的评价对象, 依据标签传播算法的基本思想, 假设同一话题下相似的句子往往共享相同的评价对象, 然后, 每个句子作为图的顶点, 句子之间的相似度由余弦相似度计算得到, 构造概率转移矩阵, 将微博文本中抽取的所有的名词短语、微博话题标签文本的分词结果作为候选评价对象, 以此构建每个节点的标签向量 (Label vector), 最后, 使用非监督标签传播算法来对候选评价对象进行排名, 从而确定最佳评价对象. 与标签传播算法不同的是, 非监督标签传播算法使用不同节点之间的链接信息来获得所有节点的正确标签.

2.4 基于条件随机场模型的方法

一些学者尝试利用解决自然语言基本任务的模型和方法来对评价对象进行抽取, 如传统的序列标注模型: 隐马尔科夫模型 (Hidden markov model, HMM)、最大熵马尔科夫模型 (Maximum entropy markov model, MEMM)、条件随机场模型 (Conditional random fields, CRFs) 等, CRF 是 HMM 和 MEMM 的改进版本, 得到了学界的一致认可并成为目前解决序列标注问题的主流方法, 限于篇幅, 关于 HMM 和 MEMM 模型, 请参考文献 [22−23]. 下面详细介绍基于 CRF 的方法.

Lafferty 等[24] 于 2001 年提出著名的 CRF 模型用于序列数据的切分和标注. CRF 模型是典型的判别式概率无向图模型. 相比 HMM 模型, 没有严格的独立性假设, 能更好捕捉上下文信息, 特征设计更加灵活. 相比 MEMM 模型, 不是局部归一化, 而是对概率进行全局归一化, 所以能避免标记偏置 (Label bias) 问题.

令 $\mathbf{x} = \{x_1, x_2, x_3, \cdots, x_n\}$ 表示观测序列, $\mathbf{y} = \{y_1, y_2, y_3, \cdots, y_n\}$ 表示状态序列, 则给定一

观测序列 $\mathbf{x}$ 的情况下 $\mathbf{y}$ 的条件概率为:

$$p(\mathbf{y} \mid \mathbf{x}, \lambda) = \frac{1}{Z(\mathbf{x})} \exp(\sum_{i=1}^{n} \sum_{j} \lambda_j f_j(y_{i-1}, y_i, \mathbf{x}, i))$$

其中, $Z(\mathbf{x}) = \sum_{\mathbf{y}} \exp(\sum_{i=1}^{n} \sum_{j} \lambda_j f_j(y_{i-1}, y_i, x, i)))$, 为归一化因子, $f_j(y_{i-1}, y_i, x, i))$ 为一特征函数, λ_j 为权值, 是模型需要学习的参数. 模型的关键在于特征函数的构造.

对于评价对象抽取问题, 则 $\mathbf{x}$ 为文本序列, $\mathbf{y}$ 为标记序列, 标注方法可采取 IOB 形式或 IOBES 形式. 若采用 IOB 形式, 则 $y_i \in \{I, O, B\}$, 若采用 IOBES 形式, 则 $y_i \in \{I, O, B, E, S\}$, 其中 B、I、E 分别表示评价对象组块的开始、内部和结束, O 表示组块外的词语, S 表示该组块只有一个词. 例如餐馆评论句子 "A few tips: skip the turnip cake and roast pork buns." 使用两种标注方法如表 4 所示.

Jakob 等[25] 利用 CRF 模型对多个领域的语料进行了实验. 选取的特征包括词特征、词性特征、依存句法特征、词距特征和观点句特征. 作者对多种特征组合进行实验发现, 在电影评论领域的准确率、召回率和 F 值最高达到 0.749、0.661 和 0.722, 但在汽车领域仅为 0.622、0.414、0.497. 其原因在于汽车领域描述评价对象的词语比电影领域更加多样, 当在测试集中遇到训练集中并未出现或出现次数非常少的评价对象词语时, 这些评价对象往往难以被识别. 这也从侧面反映了监督学习方法的局限性: 结果的好坏与训练语料的规模密切相关, 要得到好的结果往往需要大量的训练语料.

对于长句文本, 句子中通常会出现较多除评价对象之外的名词或名词短语, 这对基于词性特征的机器学习方法有较大影响. 张莉等[26] 针对此问题提出了核心句的概念 (所谓核心句, 即通过分析将长句转换为短句, 短句为原长句的关键部分, 能够代表原长句的主要观点), 并总结归纳了七种规则进行核心句的转换. 在此步骤之后, 使用基于词、词性、句法结构特征的 CRF 模型进行训练, 实验结果表明核心句方法的引入使得准确率、召回率和 F 值均有较大提升, 证明了基于核心句方法的有效性.

为了改善名词短语型评价对象的识别效果, 徐冰等[27] 提出在模型训练过程中引入浅层句法分析和启发式位置特征, 以提高评价对象识别的准确率. 所谓浅层句法分析, 是指识别文本中结构相对简单但核心的语法单位, 例如名词短语等. 浅层句法分析技术相对来说较为成熟, 主流方法在 CoNLL 2000 评测语料中 F 值可达 90 % 以上, 因此, 作者试图利用浅层句法分析器学习句法特征, 并将其加入到特征模板中, 以提高短语级别的评价对象抽取性能. 此外, 还使用了距离评价词的位置信息作为特征, 结合浅层句法分析特征一同进行训练. 在 COAE2008 任务 3 的语料进行实验, 结果显示, 启发式位置特征使评价对象识别平均 F 值提高 1.47 %, 浅层句法分析特征使平均 F 值提高 5.61 %, 两个特征同时引入时使平均 F 值提高 5.85 %.

表 5 归纳梳理了基于 CRF 模型进行评价对象抽取的方法.

2.5 基于深度学习的方法

从 2006 年以来, 深度学习在语音识别和计算机视觉等领域的重要突破重新引起了学术界的高度重视, 相比传统的机器学习方法, 深度学习可以自动地学习特征, 避免了需要大量领域知识的特征提取过程. 此外, 深度学习因有大量的可调参数和层次结构, 因而具有更强的特征表示能力.

之后, 学者开始利用深度学习自动学习特征的能力和 CRF 的优点进行融合[29−30], 深层神经网络负责学习句子文本的语义表示, 获得的分布式表示作为 CRF 模型的输入, 以此得到最终的标签. Wang 等[29] 提出使用递归神经网络 (Recursive neural networks) 和 CRF 相结合的方法进行评价对象和评价词的抽取, 在 DT-RNN(Dependency-tree RNN) 后增加一 CRF 层, 以帮助 CRF 捕捉到词语的上下文信息, 结构图如图 5 所示. 此方法虽然在实验结果上比传统的 CRF 模型并无明显优势, 但避免了手工构建特征的繁杂工作. Yin 等[30] 将学习到的嵌入 (Embedding) 串联起来作为 CRF 模型的特征, 嵌入的学习例子如图 6 所示, 其中词嵌入和依存句法上下文嵌入由 (1) 得到, 线性上下文嵌入由 (2) 得到. 也有学者完全使用深度学习的框架, Liu 等[31] 使用预先训练好的词向量初始化循环神经网络 (Recurrent neural networks) 模型的输入, 在模

表 4 IOB 和 IOBES 标注例子

Table 4 Example of IOB and IOBES annotation

	A	few	tips	:	skip	the	turnip	cake	and	roast	pork	buns	.
IOB	O	O	O	O	O	O	B	I	O	B	I	I	O
IOBES	O	O	O	O	O	O	B	E	O	B	I	E	O

表 5　基于 CRF 模型抽取方法比较

Table 5　Comparison of CRF-based extraction methods

文献	方法	级别	数据集	语言	实验结果
Jakob 等[25] 2010[1]	CRF (Lexicon-related features)	句子	Movies	English	0.749 0.661 0.702
			Web Services	English	0.722 0.526 0.609
			Cars	English	0.622 0.414 0.497
			Cameras	English	0.614 0.423 0.500
徐冰等[27] 2011	CRF (浅层句法特征 + 启发式位置特征)	句子	数码相机	Chinese	0.5097 0.3579 0.4206 (S) 0.7295 0.5122 0.6019 (L)
			手机	Chinese	0.5679 0.3631 0.4430 (S) 0.7761 0.4962 0.6054 (L)
			笔记本	Chinese	0.5843 0.4200 0.4887 (S) 0.7692 0.5529 0.6433 (L)
			汽车	Chinese	0.4302 0.2060 0.2786 (S) 0.6265 0.3001 0.4058 (L)
Liao 等[28] 2016	CRF (Lexicon-related features + Syntactic and semantic information)	句子	COAE2014 微博数据	Chinese	0.6901 0.4605 0.5523 (S) 0.7432 0.4855 0.5873 (L)

[1] 作者还进行了跨领域的实验, 这里只给出单领域的结果.

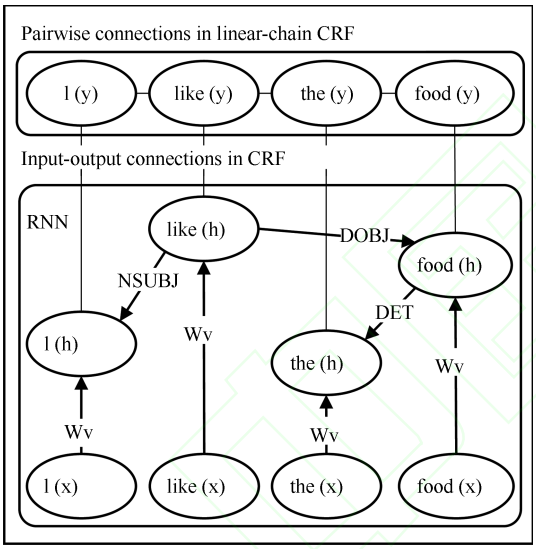

图 5　RNCRF 结构[29]

Fig. 5　Structure of RNCRF

型训练的过程中词向量得到微调. 实验结果表明, 无需任何特征工程的深度学习模型比传统的 CRF 模型得到更好的结果, 此外, 模型还可以灵活地融合词性、语块等语言学特征. Poria 等[32] 则使用了卷积神经网络 (Convolutional neural network) 识别评价对象, 利用 Amazon 评论文本训练词向量作为特征输入到 CNN 中. 表 6 归纳总结了基于深度学习进行评价对象抽取的方法.

目前深度学习在自然语言处理领域取得的进展大多来自于词语的实数向量表示, 而非采用层次结构具有抽象能力的深度学习本身[33]. 在自然语言处理领域, 此深度层次结构还没有被真正挖掘和利用. 此外, 深度学习也有其缺点, 比如需要的训练数据量比较大、计算时间复杂度较高. 相信随着计算机计算能力的提升和如今大数据的发展, 深度学习在语义表示取得关键性突破的同时也会给评价对象抽取问题带来启示.

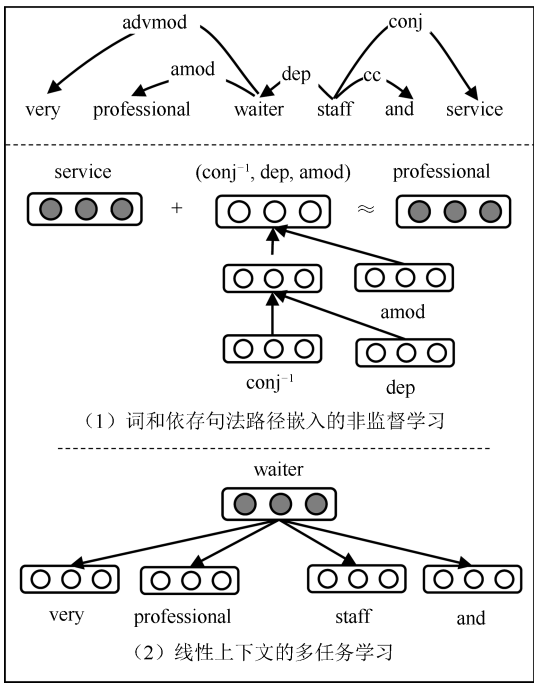

图 6　向量学习例子[30]

Fig. 6　Example of the embedding learning

表 6 基于深度学习的抽取方法比较

Table 6 Comparison of deep learning based extraction methods

文献	方法	级别	数据集	语言	实验结果
Liu 等[31] 2015	Recurrent Neural Networks + Word Embedding + Linguistic Features	句子	Laptop[1]	English	0.7457 (F1)[3] 0.7500 (F1)[4]
			Restaurant[1]	English	0.8206 (F1)[3] 0.8082 (F1)[4]
Wang 等[29] 2016	Dependency-Tree Recursive Neural Networks + CRF + Linguistic/Lexicon Features (Name List and POS Tag)	句子	Laptop[1]	English	0.7809 (F1)
			Restaurant[1]	English	0.8473 (F1)
Yin 等[30] 2016	Word Embedding + Dependency Path Embedding + CRF	句子	Laptop[1]	English	0.7516 (F1)
			Restaurant[1]	English	0.8497 (F1)
			Restaurant[2]	English	0.6973 (F1)
Poria 等[32] 2016	Convolutional Neural Network + Linguistic Patterns	句子	Laptop[1]	English	0.8672 0.7835 0.8232
			Restaurant[1]	English	0.8827 0.8610 0.8717

[1] SemEval-2014 Task 4 数据集
[2] SemEval-2015 Task 12 数据集
[3] 双向 Elman Type RNN + Amazon 词向量 + 语言学特征
[4] LSTM-RNN + Amazon 词向量 + 语言学特征

2.6 分析

目前评价对象抽取研究并没有统一的评测语料, 不同的研究者往往选取了来自不同来源 (网站)、领域、语言和语料规模的评论文本, 同一算法在不同语料中性能存在较大差异, 单纯分析最后的实验结果则不具备可比性, 因此, 本文选取了学界使用范围最广的 CRD(Customer review datasets) 和 SemEval-2014 ABSA(SemEval-2014 Aspect based sentiment analysis Datasets) 作为实验比较的基准语料, 并据此分析各种算法在同一数据集的性能表现, 由表 7 和表 8 可以看出, 在 CRD 数据集上, 基于规则的评价对象抽取方法取得了最好的 F 值, 而在 SemEval-2014 ABSA 数据集上, 则是 CRF 和深度学习获得最好结果, 笔者认为, 出现这种差异的原因在于, 在 CRD 数据集上, 数据量较少, 因此基于模板规则等工分析成分较多的方法能取得较好成绩, 而对于 SemEval-2014 ABSA 数据集而言, 数据量较大, 因此适合像 CRF 和深度学习等依赖训练数据规模的机器学习方法. 但是否这两种方法就是最好的方法了呢? 这里需要指出的是, 实验结果除了与模型方法的先进性有关之外, 还与多种因素有关, 如数据预处理阶段分词和词性标注质量、选取的情感词典差异、语法分析器优劣、模型参数调谐技巧等等, 因此仅从评测结果评判各种方法的优劣会丢失其在特定情况下的特点, 下面对各类方法特点进行深入分析.

产品评论文本中, 描述评价对象的词语有限且较快收敛, 因此基于频率的评价对象抽取方法比较常见. 文献[38] 分析指出, 在语料规模为 4 000 的书籍领域, 标注数接近 300 时便可囊括全部的评价对象. 其主要有三个方面的缺点: 首先, 只能抽取在语料中频繁出现的评价对象, 对于非频繁评价对象需要使用额外的技术手段进行抽取; 其次, 难以捕捉到

表 7 在 Customer Review Datasets 上的实验结果

Table 7 Experimental results on Customer Review Datasets

文献	方法	实验结果 (P-R-F1)
Hu et al.[3], 2004	Frequency-based	72.20 - 79.80 - 75.81
Popescu et al.[4], 2005	PMI	88.20 - 77.20 - 82.33
Qiu et al.[13], 2011	Double Propagation	88.00 - 83.60 - 85.74
Liu et al.[19], 2012	Word Alignment + Random Walk	85.80 - 86.20 - 86.00
Poria et al.[34], 2014	Rule-based	89.41 - 91.42 - 90.40
Poria et al.[32], 2016	CNN-based	90.19 - 86.18 - 88.14

表 8 在 SemEval-2014 Task 4 ABSA Datasets 上的实验结果

Table 8 Experimental results on SemEval-2014 Task 4 ABSA Datasets

文献	方法	实验结果 (P-R-F1)[1]		
Bornebusch et al.[35], 2014	Frequency-based	23.00	25.00	24.00
		37.00	40.00	38.00
Garcia-Pablos et al.[36], 2014	Pattern-based	32.10	42.50	36.60
		57.50	64.50	60.80
Chernyshevich[37], 2014	CRF-based	N/A	N/A	74.55
		N/A	N/A	79.62
Poria et al.[32], 2016	Convolutional Neural Network	86.72	78.35	82.32
	+ Linguistic Patterns	88.27	86.10	87.17

[1] 上行表示 Laptop 领域数据集实验结果, 下行表示 Restaurant 领域数据集实验结果.

句子含义, 当句子仅为陈述句表达客观事实时, 此时并无评价对象; 最后, 不相似领域间迁移效果差, 比如迁移到新闻评论领域, 因为此类型评论中频繁名词性词语中充斥了大量非评价对象.

基于模板规则的方法由于算法时间复杂度低, 无需大量标注语料, 因此往往成为工业界首选的方法. 其不足之处在于, 依赖句法分析器的质量, 而目前句法分析器在口语化严重的文本中表现差强人意; 此外, 模板规则难以适应日新月异的语言现象, 特别是当今互联网时代层出不穷的新词新意、词类活用和语法错误等现象, 不完善的规则会导致召回率下降, 错误的规则会导致准确率下降.

基于图论的方法一般通过评价对象和评价词的共现强化来实现评价对象和评价词的联合抽取, 这种思路利用了形容词和名词的先验知识, 在特定的评论句子中具有很好的效果, 如汽车产品评论" 输出输入稳定, 动力充沛, 推背感较强, 电子助力转向轻盈顺滑, 方向指向性强, 配置很高; 胎躁比较大, 底盘悬挂硬, 过减速带很颠". 其中假设评价词为形容词, 评价对象为名词或名词性短语, 然而, 这种假设具有局限性. 如" 政府官员是猪吗, 这点东西都搞不定" 和" 郎平是个铁榔头", 这两个句子均使用了隐喻的修辞手法表达了自己的情感, 但句子中并无形容词性的评价词. 另外, 此类方法存在的频繁噪音词和非频繁长尾词也是一个需要解决的问题.

基于条件随机场的方法可以较为精确地抽取出评价对象, 这取决于选取适当有效的特征. 其不足的地方主要在于, 首先, 依赖训练集的大小和标注质量, 大规模标注语料的获取非常昂贵而且语料质量参差不齐, 低质量语料对评价对象的抽取结果影响极大; 其次, 依赖于特定领域, 不同领域的语言表述方式存在极大差异, 例如产品评论和新闻事件评论, 进行领域迁移时往往需要重新训练模型.

基于深度学习的方法由于能避免大量特征工程方面的工作, 因此在最近几年受到学术界的广泛关注. 此方法的核心在于词向量, 词向量训练的时间比较久, 通常需要几个小时. 如果希望取得好的效果, 训练词向量语料通常需要是领域相关的, 因此如果需要进行领域迁移, 则需要重新训练词向量. 此外, 深度学习的方法由于具有巨大的参数空间, 通常达到百万级别, 因此需要大量的训练语料才展现出比传统方法在识别性能上的优势.

上文分析比较了各类评价对象抽取方法的特点, 下面对这些方法的优缺点进行归纳总结, 如表 9 所示.

除了方法各自特有的优缺点之外, 待处理评论文本的特征也是需要注意的问题. 新闻述评、专业影评或产品评测等书面化长文本用语相对符合正字法, 对其进行常规的自然语言处理分析 (如分词、词性标注、命名实体识别、依存句法分析等) 可以取得较好的效果, 这些分析结果在作为输入特征帮助评价对象识别中起到非常重要的作用. 而网民用户临时性评论等口语化短文本具有特征稀疏、用语不规范、噪音严重等特点, 严重影响了评价对象识别特征提取的效果. 常见的例子有拼写错误 (如 "超级丹" 写作 "超级单")、同音字 (如 "马化腾" 写作 "麻花藤")、缩写 (如 "政府" 写作 "ZF", "警察" 写作 "JC") 等, 此外, 随着微博、微信等社交媒体的流行, 人们开始习惯使用表情符号来表达自己的态度和意见. 表情符常常是充当情感增强或者表达形象化的角色, 在句子中并不构成语言单位, 移除后不影响对文本进行处理, 所以许多学者在预处理时一般不考虑这些表情符或仅作为情感标志辅助评价对象的识别. 然而, 随着表情符号的广泛使用, 越来越多的表情符开始充当句子成分, 如微博 "这个产品 [cow] 逼啊"、"真是 [pig] 一样的队友"、"这家店的 [hamburger] 挺好吃的呀呀呀", 这些表情符号的出现使得对其进行常规自然语言处理分析变得极具挑战, 在对此类文本进行评价对象识别之前需要判断这些符号是否在句子中充当了语法角色.

表 9 各种方法优缺点比较

Table 9 Comparison of advantages and disadvantages of various methods

方法	优点	缺点
基于频率的方法	1. 方法简单直观, 在产品评论领域可以取得较好效果 2. 相似领域间迁移容易	1. 抽取非频繁评价对象需要额外手段 2. 难以捕捉句子含义 3. 不相似领域间迁移效果差, 如新闻评论领域
基于模板规则的方法	1. 时间复杂度低 2. 不需要大量标注语料	1. 依赖依存句法分析器的质量 2. 编制模板规则需要专家知识
基于图论的方法	1. 结合了评价对象和情感词的共现频率关系 2. 领域可移植性强	1. 没考虑无形容词性评价词的句子 2. 频繁噪音词和非频繁长尾词需要进一步筛选
基于 CRF 的方法	1. 识别准确率高 2. 可以任意添加特征	1. 需要大量训练语料 2. 模型训练时间长 3. 领域迁移则需要重新训练
基于深度学习的方法	1. 避免大量特征工程工作 2. 识别效果好 3. 从一定程度上能从语义角度分析评价对象	1. 领域相关词向量训练耗时 2. 需要大量标注训练语料 3. 模型训练时间复杂度高 4. 领域迁移困难

对于口语化短文本, 在预处理阶段通常会进行“恢复” 操作, 如拼写检查、标点符号更正等规范化步骤, 这些规范化过程并不是简单地使用正则表达式匹配就可以解决, 有时也需要根据上下文语境进行判断, 例如 “Goood” 可能表示 “Good” 或者 “God”. 而对于短文本特征稀疏、上下文语境信息不足的问题, 主要有两种方法, 一是借助人工构建的大规模知识库 (如 Freebase) 或本体库的概念化信息进行辅助识别, 二是通过挖掘互联网海量文本中词语之间的联系 (如共现关系) 对短文本进行补充. 这两种方法均是利用了外源知识, 但前者的构建需要花费较多人工力量.

3 评价对象抽取评测及语料资源

3.1 评价标准

在评价对象抽取中, 一般采取 Precision、Recall 和 F1 作为评价标准. 计算公式如下:

$$Precision = \frac{\#SystemCorrect}{\#SystemProposed}$$

$$Recall = \frac{\#SystemCorrect}{\#Gold}$$

$$F1 = \frac{2 \times Precison \times Recall}{Precison + Recall}$$

其中, #SystemCorrect 是系统判断正确的评价对象的数目, #SystemProposed 是系统识别出来的评价对象的数目, #Gold 是人工标注的评价对象的数目 (标准答案).

通常一个评价对象由四元组 $< cid, sid, from, to >$ 界定, 其中 cid 为评论 ID, sid 为句子 ID, $from$ 和 to 分别为评价对象在句子中的起始和结束位置, 只有当系统和标准答案在四元组各个元素上均匹配时, 才判断为正确抽取.

以上评价方法是目前国内外常用的评测标准, 然而, 有学者认为此方法对于评价对象抽取来说过于严格或者对于意见挖掘的目的来说不太合适, 因此出现了基于此评价方法的一些变种, 如宽松评价方式:

首先定义覆盖率 c:

$$c(s, s') = \begin{cases} \frac{|s \cap s'|}{|s'|}, & \text{if } cid = cid' \text{ and } sid = sid' \\ 0, & \text{else} \end{cases}$$

其中, s 和 s' 分别表示系统和标准答案评价对象的起始位置区间, $\cap$ 表示计算两个区间的交集, $|\cdot|$ 表示计算区间的长度. 结果集合覆盖率 C:

$$C(S, S') = \sum_{s \in S} \sum_{s' \in S'} c(s, s')$$

其中, S 为系统结果集, S' 为标准结果集. 则最终精确率 P、召回率 R 和 F1 为:

$$P = \frac{C(S, S')}{|S|}$$

$$R = \frac{C(S, S')}{|S'|}$$

$$F1 = \frac{2PR}{P + R}$$

这种宽松评价方式在一定程度上减轻了评价对象边界争议问题, 如句子 “是义诊, 该不是议诊吧? 这么多的人, 充分说明了国家的医疗保险体系纯粹是扯淡. ”, 当系统抽取 “医疗” 作为评价对象时, 如按照严格的评测方式, 则判断为错误, 但若按照宽松评价方式, 则分数仅作了一些惩罚 (0.5). 然而, 此评价方法也存在一些弊端, 如系统抽取 “医疗保险体系” 作为评价对象时, 分数为 1, 因此当系统将整个句子作

为结果输出时可以取得非常高的分数. 由于此问题, 所以一般将宽松评价方式作为辅助的评测指标.

此外, 有学者认为, 不同的评价对象重要程度不同, 对于常见的评价对象应该赋予更高的分数. 如在产品评论中, 频率较高的评价对象往往是人们特别关注的, 而对于频率低的评价对象则往往意义不大, 可以忽略, 基于此思想, 有如下 Precision、Recall 的计算方法:

$$Precision = \frac{\sum_{i=1}^{|A|} f_i \times g(a_i, A)}{\sum_{i=1}^{|A|}}$$

$$Recall = \frac{\sum_{i=1}^{|T|} f_i \times g(a_i, T)}{\sum_{i=1}^{|T|}}$$

其中

$$g(a_i, A) = \begin{cases} 1, & a_i \in A \\ 0, & a_i \notin A \end{cases}$$

$$g(a_i, T) = \begin{cases} 1, & a_i \in T \\ 0, & a_i \notin T \end{cases}$$

A 为系统得到的评价对象集, T 为标准标注集, f_i 为评价对象 a_i 的频率. 注意这里的集合不包含位置信息, 如 size[0, 4] 和 size[10, 14] 在集合中当作一个元素, 频率为 2.

除了上述的评价方法外, 还有一些参照信息检索领域的评测指标, 如 Precision@k, 这种评价方法适用于当数据集数量太过庞大而人工难以进行标注的情况, 由于篇幅限制, 此处不再赘述, 可参考文献[39].

3.2 相关评测

目前在情感分析评测任务中包含了评价对象抽取子任务. 国际上关于评价对象抽取任务的评测主要是 SemEval. 2014 年 SemEval 任务 4“方面级情感分析 (Aspect based sentiment analysis)” 提供了 Restaurant 和 Laptop 两个领域的用户评论数据[40]. Chernyshevich[37] 实现的系统使用了词典、句法和统计特征进行训练, 在 Laptop 领域中排名第一, F 值为 74.55 %. Toh 等[41] 设计的 DLIREC 系统在 Restaurant 领域中取得最好成绩, F 值为 84.01 %, 除了词性和依存句法特征外, 还使用了来自 Yelp 和 Amazon 的外部资源进行聚类得到的词簇特征等. Brun 等[42] 的 XRCE 系统在 Restaurant 领域取得了与第一名非常接近的成绩, F 值为 83.98 %, 此系统和前两个基于 CRF 的系统不同, 其使用语义分析器抽取出 “评价词 - 评价对象” 对. 语义分析器的词典除了在训练过程中通过训练语料扩充外, 还通过维基百科和 WordNet 抽取关于餐馆和食物的词语. 2015 年 SemEval 的任务 12[43] 与 2014 年大同小异, 不同的是 2015 年不再将句子视作一个独立的单位, 而是拓展至整个评论文本. 针对评价对象抽取的评测中, 排名靠前的有 San Vicente 等[44] 的 EliXa 系统和 Toh 等[45] 的 NLANGP 系统. EliXa 系统采用基于感知器算法 (Perceptron algorithm) 进行训练, 获得了第一名的成绩, F 值达到 70.05 %. NLANGP 系统延续使用基于 CRF 的方法, 使用 CRFsuite 工具进行模型构建和训练, F 值达到 67.11 %.

国内对于评价对象抽取的评测相对较早. 在第一届中文情感倾向性分析评测 (COAE2008) 中, 要求给出每个评价句子中的评价对象并对其倾向性做出判断, 如输入句子为 “(杭州大酒店) 酒店的服务不错, 位置也不错, 可惜酒店没有游泳池”, 输出应为 “{服务, positive}, {位置, positive}, {游泳池, negative}”. Zhang 等[46] 使用基于 CRF 的模型, 取得精确评价方式第一名和宽松评价方式第二名的成绩. 刘鸿宇等[18] 设计的系统在宽松评价方式中取得第一名的成绩, 作者将提取到的名词和名词短语看作是候选评价对象, 然后使用频率过滤、PMI 和名词剪枝等算法进行筛选得到最终的评价对象. 在 2012 年 CCF 自然语言处理与中文计算会议 (NLPCC 2012) 中, 评测数据来自于腾讯微博平台用户对热门事件的评论, 如 “三亚春节宰客” 等, 评测数据包括 20 个话题, 每个话题约 1 000 条微博. 在观点句评价对象抽取任务中, 侯敏等[47] 的 CUCsas 系统取得最好成绩, 采用了基于短语情感词典及语义规则进行观点句识别和情感要素的抽取.

3.3 语料资源

除了上述评测中提供的数据外, 还有一些学者和研究机构提供了语料资源:

Customer Review Datasets

`https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets`

CRD 是 Liu 等[3] 提供的 5 个英文产品评论语料, 数据采集于 Amazon 和 CNET. 每条评论包括评论标题和评论文本主体, 均已标注了评价对象、二元情感倾向及强度.

Multi-Perspective Question Answering

`http://mpqa.cs.pitt.edu/`

MPQA 由 Wiebe 等[48] 提供, 是新闻领域意见挖掘的英语语料库, 语料来源于不同国家 187 个新闻源, 包含 535 篇新闻文档共 10 657 个句子. MPQA 语料库除了标注了评价者、评价对象和情感倾向性三元组之外, 还对新闻文本进行了言语事件、指代等深

度标注.

Darmstadt Service Review Corpus
`https://www.ukp.tu-darmstadt.de/data/sentiment-analysis/darmstadt-service-review-corpus/`
DSRC 由达姆施塔特工业大学 Ubiquitous Knowledge Processing Lab 的 Toprak 等[49] 提供, 数据来自 http://www.rateitall.com 和 http://www.epinions.com 两个在线评论网站, 包括对大学 (如凯佩拉大学) 和在线服务公司 (如 PayPal) 的评价. 此语料在句子和短语级别进行了情感信息标注. 为基于服务型机构的细粒度情感分析提供了良好的语料资源.

Arabic Opinion Target Corpus
`http://www.cs.columbia.edu/~noura/Resources.html`
AOTC 是哥伦比亚大学 Farra 等[50] 提供的阿拉伯语新闻评论语料库, 收集自卡塔尔半岛电视台新闻网站, 包含政治、文化和体育三个主题, 共 1 177 条评论, 约 7 346 个评价对象. 为研究多语言或跨语言细粒度情感分析提供了可能性.

Corpus for Implicit Aspect Extraction and Implicit Aspect Indicator Extraction:
`http://www.gelbukh.com/resources/implicit-aspect-extraction-corpus/`
此语料库由墨西哥国立理工大学 Cruz 等[51] 提供, 是 CRD 语料库的再标注, 主要标出了语料中隐式评价对象的指示词, 比如 small[size]、light[weight] 和 slick[appearance], 中括号中的词语表示隐式评价对象. 可为产品评论挖掘隐式评价对象提供思路和数据.

Chinese Product Review Opinion corpus
CPRO 是 Xu 等[52] 提供的电子产品领域中文语料库, 其中 Digital camera 包含 1 100 条评论, Mobile phone 包含 500 条评论. 总共 10 935 个句子, 7864 个句子包含观点, 12 724 个意见单位 (语料中称为 Comment). 语料中标注了评价对象、是否为隐式评价对象、评价短语、评价词、否定词、程度副词、情感极性及强度等信息, 这对产品评论文本进行深度意见挖掘非常有用.

汽车领域情感分析语料库
此语料库是 2016 CCF 大数据与计算智能大赛中 "基于视角的领域情感分析" 赛题下提供的数据, 标注了评价对象和情感极性. 语料仅限汽车领域, 而且这里的评价对象特指汽车品牌名 (语料中称为视角), 语料规模约为 8 000 个评论句子, 其中标注有 1 000 多个情感单位 (评价对象, 情感极性).

4 评价对象抽取难点

从评测结果看, 评价对象抽取离实际应用还有较大差距, 许多问题和挑战亟需解决和深入研究. 一方面在自然语言处理的任务中, 特别是在中文信息处理中, 其词义、词性、句法、语法和语义分析都有歧义问题. 另一方面, 语言是人类表达情感、交流思想、传播知识的载体, 具备很大的灵活性, 在 Web 2.0 时代更体现无遗, 网络文本充满了噪音和不规范, 网络新词、符号或旧词新意层出不穷. 此外, 评价对象抽取亦有自身特定的难题需要解决, 以下从隐式评价对象和跨句子评价对象两个角度对评价对象抽取的难点进行讨论.

4.1 隐式评价对象

隐式评价对象是文本中未出现、但根据语义可以分析出的评价对象, 此类文本往往较短, 将评价对象进行了省略表达. 短文本由于可用特征少, 使用基于机器学习的方法进行评价对象识别往往效果会非常差. 下面以具体例子进行分析:

例 1 "好是好, 就是太贵了"

例 2 "充满电才玩一个小时就没电了, 这什么垃圾手机"

例 3 "This camera will not easily fit in a pocket."[1]

例 4 "小米手机果然是为发烧而生的, 哈哈哈"

隐式评价对象往往隐含在评价词中, 例如在移动设备的评论文本中, 例 1 评价对象为 "价格", 例 2 评价对象为 "电池", 评价词 "贵" 和 "没电了" 隐含了商品评论文本所评价的对象. 隐式评价对象也可以隐含在语句描述中, 如例 3 描述了相机很难放口袋里, 评价对象为 "尺寸", 例 4 由 "发烧" 联想到 "发热", 评价对象为 "散热". 当前的隐式评价对象抽取研究往往先将评价对象事先确定好, 然后使用基于监督学习方法进行隐式评价对象分类[53][54] 或设计函数将评价词映射到隐式评价对象集合里[55]. Wang 等[53] 将隐式评价对象识别问题看作是分类问题, 使用约束的潜在狄利克里分布结合先验知识进行特征提取, 然后通过支持向量机进行隐式评价对象识别. 然而, 事先确定隐式评价对象类别的方法不仅领域适用性差, 且可能出现评价对象不属于任何预定类别的情况. 此外, 对隐式评价对象识别有时还需要常识, 如 "心疼佟丽娅, 陈思诚疑出轨" 事件中, 有网友评论 "家里有辆保时捷不开非要去挤公交", 分析该评论易知, 短短一句话却包括事件主要对象 — 男方、女方 (隐喻保时捷) 和第三方 (隐喻公交). 对隐式评价对象的有效识别需要对语句进行 "理解", 并产生原语句中并不 "存在" 的评价对象,

因此结合认知和自然语言生成技术可能是解决此问题的一个研究方向.

4.2 跨句子评价对象

在评论文本中, 经常会出现需要上下文信息才能确定其所评价的人或事物. 例如句子 “小明是一名研究生, 他不但心地善良乐于助人, 学术能力也是一流的” 中 “心地善良乐于助人” 修饰的是 “他” 这个代词, 但我们从上下文可以知道 “他” 指代的是句子首端 “小明”. 倘若我们仅抽取出 “他” 这个评价对象, 显然对于之后的情感分析等任务没有意义. 这里需要指代消解技术, 但这也是自然语言理解中的难点之一. 现有的评价对象抽取研究工作大多局限于单个句子, 对于评价信息分布在不同句子的情形尚无有效的解决方法, 将指代消解结合实体链接和篇章分析等跨句子分析技术或许是解决此问题的一个研究思路.

5 总结与展望

“仁者见仁, 智者见智”, 同一事物, 不同的人在不同时间、不同场合很可能有不同的看法和意见. 在当今互联网迅猛发展和网民意见井喷的时代, 大量的用户生成内容给我们对文本情感分析研究带来了机遇与挑战. 情感分析任务通常可以分为下面几个子任务: 评价人抽取、评价对象抽取、情感极性以及强度判别, 这些元素构成了一个完整的意见概念. 评价对象抽取作为情感分析的子任务, 在细粒度情感分析中具有举足轻重的地位. 本文在对评价对象抽取问题的研究方法、研究现状、相关评测和难点问题进行了综述. 10 多年来, 评价对象抽取在单领域、单语言语料上取得了一定的成果, 但许多相关问题的研究尚处于起步阶段, 还有许多问题亟待解决, 主要有以下几点:

1) 领域无关的评价对象抽取. 领域无关的评价对象抽取一直是评价对象抽取的难点问题. 目前的评价对象抽取几乎都限定在特定领域, 在评测中亦是如此. 而针对特定领域的评价对象抽取方法在领域迁移时往往会遇到问题, 因此寻求自动、有效、领域无关的评价对象抽取方法具有非常重要的意义. 无监督学习方法主要基于规则和语法关系, 依赖的领域知识相对较少, 因此对于解决领域无关的评价对象问题效果较好, 但泛化能力有限, 对于网络文本日新月异的表达方式和旧词新意等难以识别. 如何将有监督学习方法的泛化优点和无监督学习方法的领域无关优点结合进行评价对象抽取是当前具有挑战性的问题, 也将成为今后的研究热点.

2) 跨语言评价对象抽取. 有监督学习方法需要大量标注好的语料来训练模型以获得较高的准确率和召回率, 然而, 高质量、大规模的标注语料需要较多的人工, 且某些语言的语料严重匮乏, 不同语言之间训练语料规模的不平衡性成为制约评价对象抽取的一大障碍. Zhou 等[56] 提出了一种跨语言的解决方案, 仅利用一份已标注好的英文评论文本和一份未标注的中文评论文本, 通过 Bing Translate 和 Co-training 方法进行中文评论文本的评价对象抽取, 但效果不尽人意. 挖掘不同语言之间在主观性文本中的共性与特征转换机制成为当前细粒度情感分析的趋势.

3) 评价对象聚类分析. 网络用户评论具有更新速度快、数据量大等特点, 如何对海量评论进行准确、简洁的高质量意见聚合具有重要应用价值, 意见聚合的一个关键技术就是评价对象聚类分析. 由于实用性, 其在情感分析和意见挖掘发展之初便受到学者的重视[57−60]. 在产品评论领域, 消费者对于产品的评价主要集中在有限的几个方面, 如汽车产品的评论集中在外观、内饰、空间、配置、动力、性价比和售后等方面, 服装产品评价则主要在色彩、板型、材质、质量等方面, 常规的评价对象抽取方法会产生非常多的产品特征, 不利于消费者根据意见挖掘系统的结果对产品进行评价, 亦违背了产品评论挖掘的初衷, 因此对评价对象的聚类分析将成为重要研究内容.

基于以上分析, 未来研究工作可以围绕以下几个方面进行:

1) 基于规则的方法和基于统计的方法结合. 从评价对象抽取问题提出以来, 学界一直致力于寻求更合适的语言规则和统计模型. 基于规则的方法需要人工编写规则和模板, 需要耗费大量人力物力, 成本较高, 且系统泛化能力弱, 难以适应当前移动互联网时代涌现的各种新词和语言规则; 而基于统计的方法又难以精确地解决评价对象抽取问题. 因此, 如何将规则和统计方法进行结合, 应是未来评价对象抽取问题的一个研究方向.

2) 统一通用的计算模型. 目前评价对象抽取方法受限于特定领域, 是根据不同领域的语言特点而特别定制的方法, 因此需要设计一个在广泛领域内、语言差异大的文本中均表现良好的方法, 例如给定任意评论文本和相关背景知识, 系统能给出这段评论文本所描述的评价对象. 如何从基于评价理论的角度将评价对象问题进行统一并考虑通用的计算模型, 对于进行开放领域的评价对象抽取研究具有重要意义.

3) 海量互联网数据的有效利用. 随着互联网的发展, 用户生成内容呈指数级增长, 网络上充满了大量的未标注数据或以隐式形式存在的 “自然标注” 数据 (如由句子 “轮胎等部件” 可得知 “轮胎” 为汽

车部件), 应用深度学习方法于这些数据, 可推进评价对象抽取的研究.

总之, 评价对象抽取是细粒度情感分析中极具挑战性的难题之一, 希望本文能给进入这一领域的研究工作者带来一定的参考和启发.

References

1 Liu B. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*. San Rafael, France: Morgan & Claypool Publishers, 2012.

2 Pang B, Lee L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*. Delft, Netherlands: Now Publishers Inc, 2008.

3 Hu M Q, Liu B. Mining and summarizing customer reviews. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, USA: ACM, 2004. 168−177

4 Popescu A M, Etzioni O. Extracting product features and opinions from reviews. In: Proceedings of 2005 Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, Canada: Association for Computational Linguistics, 2005. 339−346

5 Blair-Goldensohn S, Hannan K, McDonald R, Neylon T, Reis G A, Reynar J. Building a sentiment summarizer for local service reviews. In: Proceedings of WWW 2008 Workshop on NLP in the Information Explosion Era (NLPIX 2008). Beijing, China: ACM, 2008. 339−348

6 Scaffidi C, Bierhoff K, Chang E, Felker M, Ng H, Jin C. Red Opal: product-feature scoring from reviews. In: Proceedings of the 8th ACM Conference on Electronic Commerce (ACMEC 2007). San Diego, USA: ACM, 2007. 182−191

7 Yi J, Nasukawa T, Bunescu R, Niblack W. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In: Proceedings of the 3rd IEEE International Conference on Data Mining. Melbourne, USA: IEEE, 2003. 427−434

8 Xu G, Huang C R, Wang H F. Extracting Chinese product features: representing a sequence by a set of skip-bigrams. In: Chinese Lexical Semantics. Berlin, Germany: Springer-Verlag, 2012. 72−83

9 Li S, Zhou L N, Li Y J. Improving aspect extraction by augmenting a frequency-based method with web-based similarity measures. *Information Processing & Management*, 2015, **51**(1): 58−67

10 Zhuang L, Jing F, Zhu X Y. Movie review mining and summarization. In: Proceedings of the 15th ACM international Conference on Information and Knowledge Management. Arlington, Virginia, USA: ACM, 2006. 43−50

11 Jakob N, Gurevych I. Using anaphora resolution to improve opinion target identification in movie reviews. In: Proceedings of the ACL 2010 Conference Short Papers. Uppsala, Sweden: Association for Computational Linguistics, 2010. 263−268

12 Song Xiao-Lei, Wang Su-Ge, Li Hong-Xia. Research on comment target recognition for specific domain products. *Journal of Chinese Information Processing*, 2010, **24**(1): 89−93
(宋晓雷, 王素格, 李红霞. 面向特定领域的产品评价对象自动识别研究. 中文信息学报, 2010, **24**(1): 89−93)

13 Qiu G, Liu B, Bu J, Chen C. Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 2011, **37**(1): 9−27

14 Zhang L, Liu B, Lim S H, O'Brien-Strain E. Extracting and ranking product features in opinion documents. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Beijing, China: Association for Computational Linguistics, 2010. 1462−1470

15 Zhao Yan-Yan, Qin Bing, Che Wan-Xiang, Liu Ting. Appraisal expression recognition based on syntactic path. *Journal of Software*, 2011, **22**(5): 887−898
(赵妍妍, 秦兵, 车万翔, 刘挺. 基于句法路径的情感评价单元识别. 软件学报, 2011, **22**(5): 887−898)

16 Liu Q, Gao Z Q, Liu B, Zhang Y L. Automated rule selection for aspect extraction in opinion mining. In: Proceedings of the 24th International Conference on Artificial Intelligence. Buenos Aires, Argentina: AAAI Press, 2015. 1291−1297

17 Jiang Teng-Jiao, Wan Chang-Xuan, Liu De-Xi, Liu Xi-Ping, Liao Guo-Qiong. Extracting target-opinion pairs based on semantic analysis. *Chinese Journal of Computers*, 2017, **40**(3): 617−633
(江腾蛟, 万常选, 刘德喜, 刘喜平, 廖国琼. 基于语义分析的评价对象 - 情感词对抽取. 计算机学报, 2017, **40**(3): 617−633)

18 Liu Hong-Yu, Zhao Yan-Yan, Qin Bing, Liu Ting. Comment target extraction and sentiment classification. *Journal of Chinese Information Processing*, 2010, **24**(1): 84−89, 122
(刘鸿宇, 赵妍妍, 秦兵, 刘挺. 评价对象抽取及其倾向性分析. 中文信息学报, 2010, **24**(1): 84−88, 122)

19 Liu K, Xu L H, Zhao J. Opinion target extraction using word-based translation model. In: Proceedings of 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju, Korea: Association for Computational Linguistics, 2012. 1346−1356

20 Xu L H, Liu K, Lai S W, Chen Y B, Zhao J. Mining opinion words and opinion targets in a two-stage framework. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria: Association for Computational Linguistics, 2013. 1764−1773

21 Zhou X J, Wan X J, Xiao J G. Collective opinion target extraction in Chinese microblogs. In: Proceedings of 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: Association for Computational Linguistics, 2013. 1840−1850

22 Jin W, Ho H H, Srihari R K. A novel lexicalized HMM-based learning framework for web opinion mining. In: Proceedings of the 26th Annual International Conference on Machine Learning. Montreal, Canada: ACM, 2009. 465−472

23 McCallum A, Freitag D, Pereira F C N. Maximum entropy markov models for information extraction and segmentation. In: Proceedings of the 17th International Conference on Machine Learning. Burlington, Massachusetts, USA: Morgan Kaufmann Publishers, 2000. 591−598

24 Lafferty J D, McCallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning. Burlington, Massachusetts, USA: Morgan Kaufmann Publishers, 2001. 282−289

25 Jakob N, Gurevych I. Extracting opinion targets in a single- and cross-domain setting with conditional random fields. In: Proceedings of 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge, Massachusetts, USA: Association for Computational Linguistics, 2010. 1035−1045

26 Zhang Li, Qian Ling-Fei, Xu Xin. Comment target extraction based on nuclear sentences and syntactic relations. *Journal of Chinese Information Processing*, 2011, **25**(3): 23−29
(张莉, 钱玲飞, 许鑫. 基于核心句及句法关系的评价对象抽取. 中文信息学报, 2011, **25**(3): 23−29)

27 Xu Bing, Zhao Tie-Jun, Wang Shan-Yu, Zheng De-Quan. Extraction of opinion targets based on shallow parsing features. *Acta Automatica Sinica*, 2011, **37**(10): 1241−1247
(徐冰, 赵铁军, 王山雨, 郑德权. 基于浅层句法特征的评价对象抽取研究. 自动化学报, 2011, **37**(10): 1241−1247)

28 Liao C, Feng C, Yang S, Huang H Y. A hybrid method of domain lexicon construction for opinion targets extraction using syntax and semantics. *Journal of Computer Science and Technology*, 2016, **31**(3): 595−603

29 Wang W Y, Pan S J, Dahlmeier D, Xiao X K. Recursive neural conditional random fields for aspect-based sentiment analysis. In: Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas, USA: Association for Computational Linguistics, 2016. 616−626

30 Yin Y C, Wei F R, Dong L, Xu K M, Zhang M, Zhou M. Unsupervised word and dependency path embeddings for aspect term extraction. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI-16). New York, USA: AAAI Press, 2016. 2979−2985

31 Liu P F, Joty S R, Meng H M. Fine-grained opinion mining with recurrent neural networks and word embeddings. In: Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015. 1433−1443

32 Poria S, Cambria E, Gelbukh A. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 2016, **108**: 42−49

33 Manning C D. Computational linguistics and deep learning. *Computational Linguistics*, 2015, **41**(4): 701−707

34 Poria S, Cambria E, Ku L W, Gui C, Gelbukh A. A rule-based approach to aspect extraction from product reviews. In: Proceedings of the 2nd Workshop on Natural Language Processing for Social Media (SocialNLP). Dublin, Ireland: Association for Computational Linguistics, 2014. 28−37

35 Bornebusch F, Cancino G, Diepenbeck M, Drechsler R, Djomkam S, Fanseu A N, Jalali M, Michael M, Mohsen J, Nitze M, Plump C, Soeken M, Tchambo F, Toni, Ziegler H. iTac: aspect based sentiment analysis using sentiment trees and dictionaries. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: Association for Computational Linguistics, 2014. 351−355

36 Garc í a-Pablos A, Rigau G, Cuadros M C S. V3: unsupervised generation of domain aspect terms for aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: Association for Computational Linguistics, 2014. 833−837

37 Chernyshevich M. IHS R&D Belarus: cross-domain extraction of product features using conditional random fields. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: Association for Computational Linguistics, 2014. 309−313

38 Xu Yan-Xiang, Luo Tie-Jian, Zhou Jia, Wang Zhu. Research on opinion distribution in reviews. *Journal of Chinese Information Processing*, 2014, **28**(3): 150−158
(许延祥, 罗铁坚, 周佳, 王竹. 评价文本中意见分布规律研究. 中文信息学报, 2014, **28**(3): 150−158)

39 Fang L, Liu B, Huang M L. Leveraging large data with weak supervision for joint feature and opinion word extraction. *Journal of Computer Science and Technology*, 2015, **30**(4): 903−916

40 Pontiki M, Galanis D, Pavlopoulos J, Papageorgiou H, Androutsopoulos I, Manandhar S. Semeval-2014 task 4: aspect based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: Association for Computational Linguistics, 2014. 27−35

41 Toh Z, Wang W T. Dlirec: aspect term extraction and term polarity classification system. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: Association for Computational Linguistics, 2014. 235−240

42 Brun C, Popa D N, Roux C. Xrce: hybrid classification for aspect-based sentiment analysis. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland: Association for Computational Linguistics, 2014. 838−842

43 Pontiki M, Galanis D, Papageorgiou H, Manandhar S, Androutsopoulos I. Semeval-2015 task 12: aspect based sentiment analysis. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Denver, Colorado: Association for Computational Linguistics, 2015. 486−495

44 Saralegi I S V X, Agerri R. EliXa: a modular and flexible ABSA platform. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Denver, Colorado: Association for Computational Linguistics, 2015. 748−752

45 Toh Z, Su J. NLANGP: Supervised machine learning system for aspect category classification and opinion target extraction. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Denver, Colorado: Association for Computational Linguistics, 2015. 496−501

46 Zhang S, Jia W J, Xia Y J, Meng Y, Yu H. Extracting product features and sentiments from Chinese customer reviews. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10). Valletta, Malta: European Language Resources Association (ELRA), 2010. 1142−1145

47 Hou Min, Teng Yong-Lin, Li Xue-Yan, Chen Yu-Qi, Zheng Shuang-Mei, Hou Ming-Wu, Zhou Hong-Zhao. Study on the linguistic features of the topic-oriented microblog and the strategies for its sentiment analysis. *Applied Linguistics*, 2013, (2): 135−143
(侯敏, 滕永林, 李雪燕, 陈毓麒, 郑双美, 侯明午, 周红照. 话题型微博语言特点及其情感分析策略研究. 语言文字应用, 2013, (2): 135−143)

48 Wiebe J, Wilson T, Cardie C. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 2005, **39**(2−3): 165−210

49 Toprak C, Jakob N, Gurevych I. Sentence and expression level annotation of opinions in user-generated discourse. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden: Association for Computational Linguistics, 2010. 575−584

50 Farra N, Mckeown K, Habash N. Annotating targets of opinions in Arabic using crowdsourcing. In: Proceedings of the 2nd Workshop on Arabic Natural Language Processing. Beijing, China: Association for Computational Linguistics, 2015. 89−98

51 Cruz I, Gelbukh A, Sidorov G. Implicit aspect indicator extraction for aspect-based opinion mining. *International Journal of Computational Linguistics and Applications*, 2014, **5**(2): 135−152

52 Xu R, Xia Y, Wong K F, Li W J. Opinion annotation in on-line Chinese product reviews. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08). Marrakech, Morocco: European Language Resources Association (ELRA), 2008. 1625−1632

53 Wang W, Xu H, Huang X Q. Implicit feature detection via a constrained topic model and SVM. In: Proceedings of 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: Association for Computational Linguistics, 2013. 903−907

54 Zeng L W, Li F. A classification-based approach for implicit feature identification. *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer, 2013, **8202**: 190−202

55 Su Q, Xiang K, Wang H F, Sun B, Yu S W. Using pointwise mutual information to identify implicit features in customer reviews. In: Proceedings of the 21st International Conference on Computer Processing of Oriental Languages: Beyond the Orient: the Research Challenges Ahead. Berlin, Heidelberg, Germany: Springer, 2006. 22−30

56 Zhou X J, Wan X J, Xiao J G. Cross-language opinion target extraction in review texts. In: Proceedings of the 12th International Conference on Data Mining. Brussels, Belgium: IEEE, 2012. 1200−1205

57 Zhai Z W, Liu B, Xu H, Jia P F. Grouping product features using semi-supervised learning with soft-constraints. In: Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, China: Association for Computational Linguistics, 2010. 1272−1280

58 Zhai Z W, Liu B, Xu H, Jia P F. Clustering product features for opinion mining. In: Proceedings of the 4th ACM International Conference on Web Search and Data Mining. Hong Kong, China: ACM, 2011. 347−354

59 Yang Yuan, Ma Yun-Long, Lin Hong-Fei. Clustering product features in opinion mining. *Journal of Chinese Information Processing*, 2012, **26**(3): 104−109
(杨源, 马云龙, 林鸿飞. 评论挖掘中产品属性归类问题研究. 中文信息学报, 2012, **26**(3): 104−109)

60 Zhang Y, Liu M, Xia H X. Clustering context-dependent opinion target words in Chinese product reviews. *Journal of Computer Science and Technology*, 2015, **30**(5): 1109−1119

蒋盛益 广东外语外贸大学信息科学与技术学院教授. 主要研究方向为数据挖掘和自然语言处理.
E-mail: jiangshengyi@163.com
(**JIANG Sheng-Yi** Professor at School of Information Science and Technology, Guangdong University of Foreign Studies. His research interest covers data mining and natural language processing.)

郭林东 广东外语外贸大学信息科学与技术学院硕士研究生. 主要研究方向为文本情感分析和自然语言处理. 本文通讯作者.
E-mail: guolindong1992@gmail.com
(**GUO Lin-Dong** Master student at School of Information Science and Technology, Guangdong University of Foreign Studies. His research interest covers text sentiment analysis and natural language processing. Corresponding author of this paper.)

王连喜 中山大学资讯管理学院博士研究生. 主要研究方向为数据挖掘、特征选择和自然语言处理.
E-mail: wanglianxi2012@163.com
(**WANG Lian-Xi** Ph. D. candidate at School of Information Management, Sun Yat-sen University . His research interest covers data mining, feature selection and natural language processing.)

符斯慧 广东外语外贸大学信息科学与技术学院硕士研究生. 主要研究方向为文本情感分析和自然语言处理.
E-mail: sihuifu93@outlook.com
(**FU Si-Hui** Master student at School of Information Science and Technology, Guangdong University of Foreign Studies. Her research interest covers text sentiment analysis and natural language processing.)