DOI: 10.7631/issn.1000-2243.2013.03.0297 文章编号: 1000-2243(2013)03-0297-08

结合领域知识的中文句子评价对象抽取

雷志城，廖祥文

(福州大学数学与计算机科学学院，福建 福州 350116)

摘要: 针对评价对象存在领域相关性这一特点，在条件随机场模型中结合领域词词典特征进行中文句子评价对象的抽取，然后利用领域规则对抽取结果进行处理. 针对COAE2011任务三标注语料的抽取实验结果表明，结合领域词词典和领域规则对于利用线性链、跳跃链和层叠条件随机场模型的中文句子评价对象抽取方法可以有效地提高抽取的精度，并抽取出更多的评价对象.

关键词: 中文句子; 评价对象抽取; 领域知识; 领域词词典

中图分类号: TP391 **文献标识码**: A

Integrate domain knowledge to extract opinion targets from Chinese sentences

LEI Zhi-cheng, LIAO Xiang-wen

(College of Mathematics and Computer Science, Fuzhou University, Fuzhou, Fujian 350116, China)

Abstract: In this paper we focus on the feature that opinion targets are domain related, and employ conditional random fields (CRFs) to extract the opinion targets incorporating domain dictionary, then we use domain rules in post processing. The method is evaluated on the labeled corpus of task 3 in COAE2011. The experiment results show that when employing the model of linear-chain, skip-chain and cascaded CRFs to extract opinion targets from Chinese sentences, it is helpful to integrate both the domain dictionary and the rules. Our method can effectively improve the accuracy and extracts more objects simultaneously.

Keywords: Chinese sentences; opinion targets extraction; domain knowledge; domain dictionary

0 引言

随着互联网的迅猛发展，越来越多的网民通过论坛、评论、微博等网络媒介来表达自己的情感、观点及看法. 文本倾向性分析，是对说话人的态度(或称观点、情感)进行分析，也就是对文本中的主观性信息进行分析[1].

评价对象抽取指在句子中抽取评论所针对的对象或对象的属性，是文本倾向性分析的重要组成部分. 它能够应用在许多相关领域，如电子商城通过研究买家对商品的评论，生成总结报告，为更好地做决策提供支持; 普通的买家，在购物时希望通过浏览其他买家的评论找得更物美价廉的商品. 但是目前网上的评论数量非常庞大，且以极快的速度在增长中，由人工完成工作量巨大且难免受个人主观影响，这就需要自动地抽取出评论所针对的对象或对象的属性以便进一步进行信息的挖掘.

迄今为止，评价对象抽取已经吸取了国内外学者的广泛关注，并开展了很多相关的研究和评测工作，NTCIR-7、NTCIR-8(NII test collection for IR systems)连续两届设立了抽取评价人和评价对象的MOAT任务，中文倾向性分析评测COAE 2008、2009、2011连续三届开展了中文观点倾向性相关要素的抽取任务.

目前的研究工作已经取得了很多成果，但是评价对象本身存在领域相关性，即对于不同领域，所使用的名词等词汇差异很大，构成评价对象的词汇差异很大，这给专业词汇的评价对象抽取带来一定困难. 目

收稿日期: 2012-05-18

通讯作者: 廖祥文(1980-)，讲师，E-mail: liaoxw@fzu.edu.cn

基金项目: 福建省自然科学基金资助项目(2010J05133); 福建省科技创新平台计划资助项目(2009J1007)

前大多数研究工作在采用机器学习方法时尚未充分利用领域知识，导致未能很好地抽取出专业名词的评价对象，抽取精度仍有待提高. 为此，本研究提出结合领域知识的中文句子评价对象抽取方法，利用线性链、跳跃链和层叠条件随机场模型结合领域词词典特征进行抽取，然后运用基于领域知识的规则进行处理，有效抽取出更多更准确的评价对象，提高了抽取的准确率和召回率.

1 相关工作

针对评价对象抽取已经开展了很多研究工作，其中 Hu 等[2-3]针对网上评论采用基于关联规则的方法. 关联规则主要由领域专家归纳总结，一般易于理解，但是无法保证规则库的系统和完备，而且对于不同的领域，适用的规则根据领域的特点也不尽相同，造成该方法的移植性有所欠缺.

此外，文献[4-5]对语料进行语法分析，文献[6]分析句法路径以识别句子的情感评价单元，文献[7]通过对句子的句法分析和依存分析进行评价对象的抽取，而在文献[8]中，Kim 等人采用语义角色标注(semantic role labeling) . 这些均属于自然语言处理(natural language processing，NLP) 的方法，通过语法、语义分析解析句子的构成，识别评价对象，适用于句子成分完整、语义清晰的文本，但处理句子成分缺失或者长度偏长的文本存在一定困难.

此外，一种很重要的方法是机器学习方法. 它通过建立统计模型进行评价对象的识别，按照模型的自动化程度可以分为非/半监督和监督两种. 非/半监督的方法通过自举、聚类、繁殖等方式实现评价对象的抽取. Jin 等[9]在自举之后通过 Lexical - HMM 模型进行抽取; 宋晓雷等[10]在抽取中运用了模糊匹配、自举和 K - means 聚类等方法; Qiu 等[11]采用双向传播扩展评价词，抽取评价对象; Li Zhuang 等[12]综合 WordNet、统计数据和先验知识进行抽取; 李娟等[13]采用基于模板的方法对中文人物评论语句进行意见元素挖掘; Xia 等[14]提出一种意见目标网络模型. 非/半监督的机器学习无需人工标注大量的训练语料，但准确率有待提高.

监督的方法通过对训练语料学习得到模型参数，然后进行抽取，它需要进行语料的人工标注，准确率高. 文献[15-17]中采用了最大熵模型，Ma 等[18]运用 centering theory 和全局信息进行抽取，除此之外，经常采用的一种模型是 CRFs(conditional random fields，CRFs) [19]条件随机场，它是一种基于无向图的判别模型，避免了最大熵马尔可夫模型(MEMM) 等存在的标记偏置问题，同时模型可以灵活地引入各种特征，在文本处理任务中有较好的表现. 文献[20-23]和 COAE 测评的多支队伍在评价对象的抽取中均采用了线性链条件随机场模型，并取得了不错的结果.

条件随机场模型在评价对象抽取方面有着不错的表现，但是，评价对象抽取存在一个重要的问题就是评价对象的领域相关性问题，对于不同的领域，所使用的专业词汇差异较大，构成评价对象的词汇差异较大，造成 CRFs 模型无法非常有效地进行专业性较强的评价对象抽取. 因此本研究在利用线性链、跳跃链和层叠条件随机场模型，并使用词、词性、语法依赖和最近名词等特征的基础上结合领域词典特征进行抽取，以更有效地抽取出评价对象. 然后对抽取结果进行基于领域规则的处理，进一步提高评价对象抽取结果的精度.

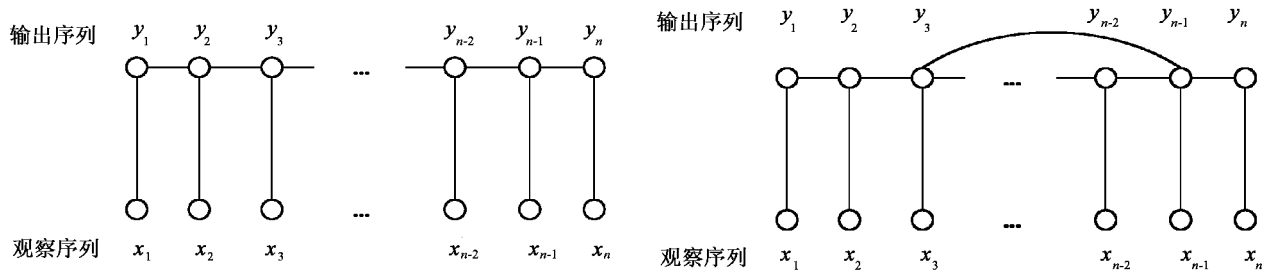
2 结合领域知识的评价对象抽取方法

2.1 单层条件随机场模型

条件随机场模型[19]一种是基于无向图的随机场模型，已经在观点挖掘和信息抽取的很多任务中被采用，并取得了不错的成果. 线性链条件随机场是由单一线性链方式构成的条件随机场模型，其模型如图 1(a) 所示.

线性链条件随机场模型可以较好地完成序列化标注任务，在评价对象抽取中已经被广泛使用并有着不错的结果，但是它无法有效解决句子中的长距离依赖问题，例如“NOKIA1110 很 结实，我 很 喜欢 1110. ”这个句子中前后两个分句中的“NOKIA1110”和“1110”指代的都是“NOKIA 1110 手机”，应该带有相同的评价对象标签. 线性链条件随机场模型无法很好地解决这一问题，为此本研究采用跳跃链条件随机场模型[24]，在线性链模型的基础上，对于可能存在长距离依赖的词，如“NOKIA1110”与“1110”之间建

立跳跃链. 模型的结构如图 1(b) 所示.



(a) 线性链　　(b) 跳跃链

图 1　线性链和跳跃链条件随机场模型的无向图结构

Fig. 1　Graphical structures of the linear – chain and the skip – chain CRFs

给定输入序列 $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$，则线性链和跳跃链模型的输出序列 $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$ 的定义分别为:

线性链: $$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})}\prod_{t}^{T}\Psi_t(y_t, y_{t-1}, \boldsymbol{x})$$

跳跃链: $$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})}\prod_{t}^{T}\Psi_t(y_t, y_{t-1}, \boldsymbol{x})\prod_{(u,v)\in\boldsymbol{\Gamma}}\Psi_{uv}(y_u, y_v, \boldsymbol{x})$$

其中: $Z(\boldsymbol{x})$ 是各自对应的归一化因子，Ψ_t 是线性链的势函数，Ψ_t 定义如下:

线性链: $$\Psi_t(y_t, y_{t-1}, \boldsymbol{x}) = \exp\left\{\sum_{k}\lambda_k f_k(y_t, y_{t-1}, \boldsymbol{x}, t)\right\}$$

$f_k(y_t, y_{t-1}, \boldsymbol{x}, t)$ 是线性链模型的特征函数，λ_k 是其对应的权重，而在跳跃链模型中，除了线性链，还有跳跃链存在，因此以 $\boldsymbol{\Gamma}$ 指代所有跳跃链$(u - v)$ 的集合，以 Ψ_{uv} 表示跳跃链对应的势函数，则跳跃链条件随机场模型的 Ψ_t，Ψ_{uv} 定义如下:

跳跃链: $$\Psi_t(y_t, y_{t-1}, \boldsymbol{x}) = \exp\left\{\sum_{k1}\lambda_{k1} f_{k1}(y_t, y_{t-1}, \boldsymbol{x}, t)\right\}$$

$$\Psi_{uv}(y_u, y_v, \boldsymbol{x}) = \exp\left\{\sum_{k2}\lambda_{k2} f_{k2}(y_u, y_v, u, v, \boldsymbol{x})\right\}$$

$f_{k1}(y_t, y_{t-1}, \boldsymbol{x}, t)$ 和 $f_{k2}(y_u, y_v, u, v, \boldsymbol{x})$ 分别指代的是跳跃链模型中的线性链和跳跃链对应的特征函数，$\boldsymbol{\theta}_1 = \{\lambda_{k1}\}_{k1=1}^{K1}$ 是线性链对应的特征函数的权重参数集，$\boldsymbol{\theta}_2 = \{\lambda_{k2}\}_{k2=1}^{K2}$ 是跳跃链对应的特征函数的权重参数集，$\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$ 是跳跃链模型总的参数集，$\boldsymbol{\theta}$ 通过对标注语料的学习得到.

2.2　层叠条件随机场模型

层叠条件随机场模型(cascaded CRFs，CCRFs) [25-28] 是在单层模型的基础上按层叠加建立起的多个层次(含两层) 的条件随机场模型，各层模型之间呈线性组合关系. 通过低层模型识别出初步结果，进行过滤和整合，处理初步结果中存在的复合词识别错误、未登录词等情况，然后将处理后的识别结果输入到高层，为高层条件随机场提供决策支持. 层叠模型如图 2 所示.

低层模型中，给定观察序列 $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ 的前提下，通过线性链模型得到 $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$，表示候选评价对象的序列.

对于候选评价对象序列 $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$ 存在的一些错误经过中间层模型的处理后转变为 $\boldsymbol{y}^1 = (y_1^1, y_2^1, \cdots, y_n^1)$，然后结合其他输入变量得到高

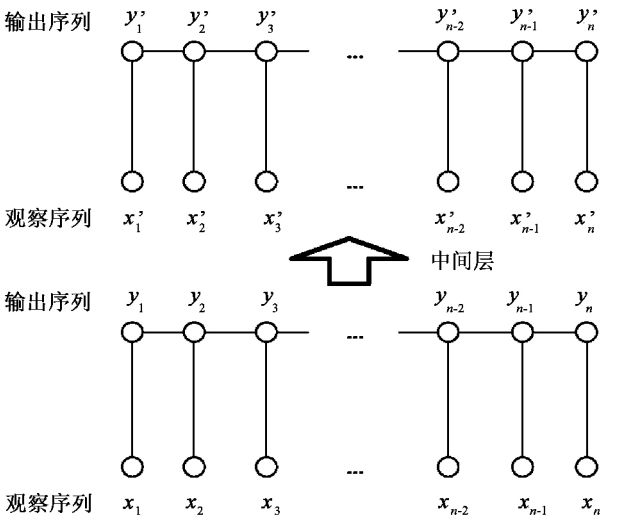


图 2　层叠条件随机场模型的无向图结构

Fig. 2　Graphical structures of the cascaded CRFs

层 CRFs 模型的观察序列 $x' = (x_1', x_2', \cdots, x_n')$ ，最终经过高层线性链模型得到 $y' = (y_1', y_2', \cdots, y_n')$ ，即评价对象的标记序列.

2.3 领域词词典

对于中文句子的评价对象抽取，融合词、词串、语法依赖和最近名词特征可以较为有效地抽取出评价对象，但是，由于词的领域相关性比较大，评价对象的领域相关性也比较大，比如电子产品名在电子领域比较可能是评价对象，而在电影领域，电影或者演员、导演名比较可能成为评价对象. 因此，针对各个领域的特点，针对语料中 3 个不同领域(电子、娱乐、经济) 分别建立领域词词典. 领域词词典的构成如表 1 所示.

表 1 领域词词典

Tab. 1 The domain dictionary

领域	组成	词数	示例
电子(dic_it)	1、品牌 2、专业词汇 3、产品属性 4、型号 5、其他	512	1、诺基亚 2、分辨率 3、显示屏 4、N96
娱乐(dic_en)	1、电影 2、演员、导演名 3、专业词汇 4、其他	250	1、阿甘正传 2、张艺谋 3、舞台
经济(dic_eco)	1、专业词汇 2、股票相关 3、其他	126	1、期货 2、股票 3、指数

2.4 特征选择

条件随机场模型可以灵活地定义各种特征，本研究前期就特征对于评价对象抽取的影响开展了一些研究，实验结果表明在线性链条件随机场中使用词、词性、语法依赖、最近名词和句子倾向性的特征可以有效地抽取出中文句子的评价对象. 在本研究中，由于所处理的句子均带倾向性，采用的是词、词性、语法依赖关系、最近名词特征. 此外针对评价对象的领域相关性，本研究结合领域词词典特征. 在线性链、跳跃链和层叠条件随机场模型中均采用了这些特征.

1) 词特征: 表示当前的词串信息;

2) 词性特征: 表示当前词串的词性，词性对于鉴别一个词是否评价对象是很重要的信息，对于多义词和不常用词的处理帮助更大;

3) 语法依赖特征: 表示当前词是否是与评价短语存在依赖的代词/名词短语，评价对象与评价短语之间在很多情况下存在着语法依赖关系，因此通过标记出当前词与评价短语是否依赖，有助于识别这些评价对象. 本研究的实验采用 Stanford parser 工具进行语法依赖的分析，评价短语则直接采用语料标注出的评价短语;

4) 最近名词特征: 表示当前词是否是离评价短语最近的代词/名词短语. 并非每个评价对象都与评价短语之间存在着语法依赖，因此，本研究识别出离评价短语最近的代词/名词短语并将此作为一个特征;

5) 领域词词典特征: 表示当前词是否包含于领域词词典中，领域词词典有助于识别专业名词评价对象. 当其他特征不足以判别评价对象与否时，通过判别是否领域专有词，能够得出评价对象的标签. 实验中采用 dic_it 表示该词属于电子领域领域词，dic_en 表示娱乐领域，dic_eco 表示经济领域.

在线性链、跳跃链和层叠条件随机场模型中均采用 [-2，2] 特征窗口. 在层叠随机场模型中对于低层模型采取以上 5 个特征组合识别出候选评价对象，然后对候选评价对象采用了中间层模型进行过滤和整合，并将处理后的候选评价对象集作为一个特征输入到高层模型中，在高层模型中采用的是以上 5 种特征组合 + 候选评价对象识别出最终的评价对象. 对于跳跃链模型中本研究在相同或相似的代词/名词短语建立跳跃链，使用的也是以上 4 种特征.

2.5 领域规则

通过条件随机场模型抽取出评价对象后，存在着一些明显的错误，例如将标点符号标识为评价对象等. 而且针对不同领域，出现的错误也有区别. 例如: 在电子领域，经常将数词标识为评价对象; 在娱乐领域，抽取出来的电影名只包含半个书名号; 在经济领域，由于句子较长，经常标识出多个名词作为评价对象等.

因此，针对这些特点，根据领域的不同采取领域知识的规则进行处理，所使用的规则如下:

首先，对于句子中未标识出任何评价对象的情况，根据电子、娱乐、经济领域各自不同的特点采用以下规则:

1) 针对电子领域，搜索句子中是否存在如“机身”，“性价比”等专业词汇，若存在则依据这些词的语法依赖特征和最近名词特征判断是否评价对象，即如果存在语法依赖或是最近名词则判定为评价对象，如果语法依赖特征和最近名词特征均无效，但这些词位于句子的开头或结尾则将其标识为评价对象，否则不作处理;

2) 针对娱乐领域，查找句子中是否存在电影名(重复的不计算) ，如果有且仅有一电影名(含书名号) ，且该电影名在句子中出现多次，则判定其为评价对象，否则依据该电影名在语料中出现的频率是否超过阈值判断是否评价对象; 如果出现多个电影名(个数 >1) ，查找与评价短语存在依赖的名词或者最近名词是否存在人名，若符合条件的人名有且仅有一个，标识为评价对象，若不存在或存在多个，不作处理;

3) 针对经济领域，查找句子中是否存在“NUM + 点”，“指数”等股票词汇，若存在，则搜索股市等相关词汇，对于出现次数超过 2 次的名词专业词汇标识为评价对象;

然后，在经过以上规则的处理后，对存在的噪声进行如下过滤:

1) 若抽取出的评价对象含标点符号(非书名号或引号) ，如“;”“.”等，且该标点符号单独出现(非数字中的小数点等) ，则判定该评价对象是噪声，进行过滤;

2) 对于电影领域，如果评价对象中包含书名号或引号，但不成对，判断出抽取有错，然后搜索句子中是否包含未抽取出的缺失的书名号或引号，若不包含则将该评价对象过滤，否则将评价对象补充完整;

3) 若评价对象为“的”，“地”等词汇或包含“而且”“虽然”等连词，判定其为错误的评价对象进行过滤;

4) 对于抽取出的评价对象，如果该评价对象属于领域词词典的某个词的子串，但不属于领域词词典，判断该词是因为分词错误而导致抽取错误的一些词，如“性价”(性价比) 搜索对应的句子中是否有未抽取出的子串，如果有则根据语法依赖或最近名词特征进行删除或者补全，如果没有，判定抽取有错，进行过滤;

5) 对于电子领域中出现数词作为评价对象的情况，搜索该数词是否多次出现，若在句子出现有且仅有一次且该词不位于句首也不位于句尾，判定为错误的评价对象，过滤.

3　实验结果与分析

3.1　语料

采用第三届中文倾向性分析评测(COAE2011) 任务 3 评价搭配抽取标注语料中所有人工标注的句子作为实验语料，每个句子含 0 至 4 个评价搭配(评价对象 + 评价短语 + 评价倾向性) . 语料的具体情况如表 2 所示.

表 2　分领域语料统计表

Tab. 2　The detail of corpus

领域	句子总数	评价对象个数
电子	5 715	7 159
娱乐	1 224	1 316
经济	513	577
总计	7 452	9 052

3.2　实验结果与分析

使用 Mallet 工具包的 GRMM 扩展包来实现线性链、跳跃链和条件随机场模型，实验中为了减少人为因素的影响，采取 4 倍交叉验证的方式. 评价短语采用标注语料中的评价短语. 进行的对比实验包括: LC_CRF(线性链) ，SK_CRF(跳跃链) ，CCRF(层叠条件随机场) ，在未加入领域词词典特征时采用的是词 + 词性 + 语法依赖 + 最近名词特征组合，加入后使用的是词 + 词性 + 语法依赖 + 最近名词 + 领域词词典特征组合，对比实验结果如表 3，表 4 所示. 通过对表 3 和表 4 的观察可以看出:

1) 跳跃链模型相对于线性链模型准确率提升显著，在 3 个不同领域约有 0.7% ~7% 的提升，另一方面，召回率也有不同程度的提升，总体 F1 值提升效果显著.

跳跃链的引入可以有效应对长距离依赖问题，例如“虽然 诺基亚 1110 有 点 老，不是 智能 手机 ，但是 它 很 耐用 . ”这个句子中“诺基亚 1110”，“智能手机”均可能是评价对象，因为“诺基亚 1110”与“它”之间的跳跃链关系，判断“诺基来 1110”为评价对象. 因此跳跃链模型可以抽取出更多准确的评价对象，达到预期的结果.

表 3 评价对象抽取的对比实验结果

Tab. 3 The experimental results of opinion targets extraction (%)

DATA	Model	未加入领域词典特征			加入领域词典		
		Precision	Recall	F1_measure	Precision	Recall	F1_measure
电子	LC_CRF	57.20	45.72	50.82	57.03	46.26	51.09
	SK_CRF	60.36	45.50	51.89	60.21	47.05	52.83
	CCRF	59.71	48.69	53.64	58.72	49.74	53.86
娱乐	LC_CRF	55.64	43.09	48.57	58.05	46.26	51.49
	SK_CRF	57.87	44.57	50.36	59.00	48.82	53.43
	CCRF	64.32	57.75	60.86	65.82	58.57	61.99
经济	LC_CRF	54.07	38.21	44.78	55.41	39.49	46.11
	SK_CRF	54.74	42.62	47.92	57.86	45.02	50.64
	CCRF	65.10	52.17	57.92	66.36	53.52	59.25

表 4 领域规则处理后的对比实验结果

Tab. 4 The experimental results of opinion targets extraction after postprocessing (%)

DATA	Model	未加入领域词典特征			加入领域词典		
		Precision	Recall	F1_measure	Precision	Recall	F1_measure
电子	LC_CRF	60.22	45.96	52.14	60.03	46.41	52.35
	SK_CRF	63.16	45.85	53.13	62.79	47.35	53.99
	CCRF	61.81	49.08	54.71	60.95	50.02	54.95
娱乐	LC_CRF	60.54	43.02	50.30	62.81	46.26	53.28
	SK_CRF	64.96	44.57	52.87	64.54	48.82	55.59
	CCRF	66.91	57.67	61.95	68.64	58.57	63.21
经济	LC_CRF	60.88	38.44	47.12	62.77	39.71	48.64
	SK_CRF	61.78	42.82	50.58	65.42	45.20	53.46
	CCRF	68.79	52.31	59.43	69.73	53.66	60.65

2) 层叠模型相对单层线性链模型不论准确率还是召回率提升都很显著，召回率在 3 个领域中表现都最好，而且总体上电子、娱乐、经济领域的 F1 值分别有大约 2.5%、10%、12% 的提升，在这 3 个领域中的 F1 值都最好. 层叠模型能够有效识别复合名词的评价对象，对于抽取出更多的评价对象，提升抽取精度是有效的.

3) 通过表 3 和表 4 的加入词典特征的前后对比可以看出，加入领域词典特征后，对于不同领域不同模型的召回率均有上升，可以识别出更多的评价对象，虽然电子领域的准确率略有下降，但总体的 F1 值有上升.

例如 “全键盘商务手机诺基亚 E71(改版机)，以超薄的时尚外形以及全面的商务功能使它成为了诺基亚商务手机里的明星”这个句子的正确评价对象是“诺基亚 E71”，评价短语是“诺基亚商务手机里的明星”，在未采用词典时因为“它”与评价短语最近，识别结果为“它”，在引入领域词典之后，因为“诺基亚”是领域词，识别结果为“它”和“诺基亚 E71”，虽然未能排除“它”，但正确识别出了“诺基亚 E71”，领域词典的引入如预期可以帮助条件随机场模型特别是跳跃链随机场模型抽取出更多评价对象.

4) 通过表 3 和表 4 的引入领域规则处理的前后对比可以看出，针对不同领域准确率提升约 1.5% ~ 3%，F1 值约提升 1.0% ~1.5%，领域规则能够针对不同领域评价对象集中的错误进行有效的过滤，对于电子、娱乐经济领域的准确率的提升显著.

例如 “显卡的话性价比一般般，适合那些对显卡要求不是很高的人群骸!”这个句子中“显卡”和“性价比”均是领域词，在引入领域规则处理前识别为空，符合电子领域规则第 1 条，“性价比”离评价短语“一般般”最近，判定其为评价对象，而显卡虽为领域词，但条件不符合，因此能够正确甄别. 领域规则在

对召回率影响不大的基础上，有效地提高了抽取的精度.

5) 在使用词典特征和领域规则之后，对于 3 个领域，相比原来的方法线性链模型 F1 值分别有 1.55%、4.71%、3.68% 的提升；跳跃链模型则分别有 2.1%、5.21%、5.54% 的提升；层叠模型则分别有 1.31%、2.35%、2.73% 的提升. 虽然对于不同领域提升有所差异，但是总体上领域知识的引入使条件随机场模型可以更加有效地抽取出中文句子的评价对象.

综上所述，领域词词典和领域规则可以有效地提高抽取的准确率和召回率. 结合领域知识，采用线性链、跳跃链和层叠条件随机场模型的方法可以抽取出更多的评价对象，并提高抽取的精度，能够很好地应用于中文句子评价对象抽取任务.

3.3　部分错误举例与分析

1) 长复合词评价对象的错误识别与分析.

句子：摩托罗拉 A3000 的背部设计科技感十足.

识别结果：空.

正确结果：摩托罗拉 A3000 的背部设计.

原因：评价对象含多个词，可分词为“摩托罗拉/nz A3000/x 的/ude1 背部/n 设计/vn”，“摩托罗拉”虽属于领域词词典，但后缀太长，且句子中不存在长距离依赖，对于跳跃链和层叠模型均由于特征不足，导致漏识别.

2) 短评价对象的错误识别与分析.

句子：而 688 不负众望，也使用了 9203624 的中央处理器，有了一颗强大的“芯”.

识别结果：空.

正确识别结果：688.

原因：688 是一型号，虽然离“不负众望”这一评价短语较近，但因为判别为数词未能使用语法依赖和最近名词上下文特征，不属于领域词词典，符合领域规判定条件，对 688 进行过滤，因此模型漏识别.

3) 非名词短语的错误识别与分析.

句子：透过流着雨水的窗户看风景也并不愉快.

识别结果：风景.

正确识别结果：看风景.

原因　“看/v 风景/n”是一动名词，风景与评价短语存在语法依赖，且风景为名词，动名词作为评价对象的频度低，名词作为评价对象的频度高，条件随机场模型错误识别，无有效领域规则进行处理.

4　结语

针对评价对象领域相关的特点，提出结合领域知识的评价对象抽取方法. 采用线性链、跳跃链和层叠条件随机场模型，并在使用词、词性、语法依赖和最近名词特征的基础上结合领域词典特征，抽取出更多准确的评价对象. 然后对于抽取的结果进行基于领域规则的处理，进一步优化了抽取结果，提高了评价对象抽取的精度. 目前工作对于抽取非名词短语及一些长的评价对象仍存在不足，将是未来的一个研究方向；同时，针对领域知识带来的移植性问题，将更多地结合领域自适应等技术提高模型的泛化能力.

致谢：感谢中国科学院计算技术研究所为本研究提供 ICTCLAS 分词工具.

参考文献：

[1] 黄萱菁，赵军. 中文文本情感倾向性分析 [J]. 中国计算机学会通讯，2008，4(2)：41 – 46.

[2] Hu Min – qing，Liu Bing. Mining opinion features in customer reviews [C] //Proceedings of Nineteenth National Conference on Artificial Intelligence (AAAI – 2004). California：[s. n.]，2004：755 – 760.

[3] Hu Min – qing，Liu Bing. Mining and summarizing customer reviews [C] //Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle：[s. n.]，2004：168 – 177.

[4] 刘鸿宇，赵妍妍，秦兵，等. 评价对象抽取及其倾向性分析 [J]. 中文信息学报，2010，24(1)：84 – 88.

[5] Lu Bin. Identifying opinion holders and targets with dependency parser in Chinese news texts [C] //Proceedings of the NAACL HLT 2010 Student Research Workshop. Los Angeles: [s. n.], 2010: 46 – 51.
[6] 赵妍妍，秦兵，车万翔，等. 基于句法路径的情感评价单元识别 [J]. 软件学报，2011，22(5)：887 – 898.
[7] 王卫平，孟翠翠. 基于句法分析与依存分析的评价对象抽取 [J]. 计算机系统应用，2011，28(8)：52 – 57.
[8] Kim S M，Hovy E. Extracting opinions，opinion holders，and topics expressed in online news media text [C] //Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text. Sydney: [s. n.]，2006: 1 – 8.
[9] Jin Wei，Hung H H，Srihari K R. Opinion miner: a novel machine learning system for Web opinion mining and extraction [C] //Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris: [s. n.]，2009: 1 195 – 1 204.
[10] 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 [J]. 中文信息学报，2010，24(1)：89 – 93.
[11] Qiu Guang，Liu Bing，Bu Jia – jun，*et al*. Opinion word expansion and target extraction through double propagation [J]. Computational Linguistics，2011，37(1)：9 – 27.
[12] Zhuang Li，Jing Feng，Zhu Xiao – yan. Movie review mining and summarization [C] //Proceedings of the ACM 15th Conference on Information and Knowledge Management. Arlington: [s. n.]，2006: 43 – 50.
[13] 李娟，张全，贾宁，等. 基于模板的中文人物评论意见挖掘 [J]. 计算机应用研究，2011，27(3)：833 – 836.
[14] Xia Yun – qing，Hao Bo – yi，Dai Liu – ling. Term extraction from Web reviews with opinion heuristics [C] //Proceedings of the Eighth International Conference on Machine Learning and Cybernetics. Baoding: [s. n.]，2009: 3 516 – 3 521.
[15] Kim S M，Hovy E. Identifying opinion holders for question answering in opinion texts [C] //Proceedings of AAAI – 05 Workshop on Question Answering in Restricted Domains. Pennsylvania: [s. n.]，2005: 20 – 26.
[16] 章剑锋，张奇，吴立德，等. 点挖掘中的主观性关系抽取 [J]. 中文信息学报，2008，22(2)：5 – 59.
[17] 方明，刘培玉. 基于最大熵模型的评价搭配识别 [J]. 计算机应用研究，2011，28(10)：3 714 – 3 716.
[18] Ma Teng – fei，Wan Xiao – jun. Opinion target extraction in Chinese news comments [C] //Proceedings of the 23rd International Conference on Computational Linguistics: Poster Volume. Beijing: [s. n.]，2010: 782 – 790.
[19] Lafferty J D，McCallum A，Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C] //Proceedings of the 18th International Conference on Machine Learning. Williamstown: [s. n.]，2001: 282 – 289.
[20] Jakob N，Gurevych I. Extracting opinion targets in a single – and cross – domain setting with conditional random fields [C] //Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Vancouver: [s. n.]，2010: 1 035 – 1 045.
[21] 徐冰，赵铁军，王山雨，等. 基于浅层句法特征的评价对象抽取研究 [J]. 自动化学报，2011，37(10)：1 241 – 1 247.
[22] 张莉，钱玲飞，许鑫. 基于核心句及句法关系的评价对象抽取 [J]. 中文信息学报，2011，25(3)：25 – 32.
[23] Ding Sheng – chun，Jiang Ting. Comment target extraction based on conditional random field & domain ontology [C] //Proceedings of 2010 International Conference on Asian Language. Harbin: [s. n.]，2010: 189 – 192.
[24] Sutton C，McCallum A. An introduction to conditional random fields for relational learning [J]. Introduction to Statistical Relational Learning，2007: 93 – 127.
[25] 刘康，赵军. 基于层叠 CRFs 模型的句子褒贬度分析研究 [J]. 中文信息学报，2008，22(1)：123 – 128.
[26] 周俊生，戴新宇，尹存燕，等. 基于层叠条件随机场模型的中文机构名自动识别 [J]. 电子学报，2006，34(5)：804 – 809.
[27] 杨晓东，晏立，尤慧丽. CCRF 与规则相结合的中文机构名识别 [J]. 计算机工程，2011，37(8)：169 – 174.
[28] 郭剑毅，薛征山，余正涛，等. 基于层叠条件随机场的旅游领域命名实体识别 [J]. 中文信息学报，2009，23(5)：47 – 52.

(责任编辑：沈芸)