文章编号：1003-0077(2013)03-0069-08

基于层叠 CRFs 的中文句子评价对象抽取

郑敏洁[1]，雷志城[2]，廖祥文[2]，陈国龙[2]

（1. 福州大学 物理与信息工程学院，福建 福州 350108；
2. 福州大学 数学与计算机科学学院，福建 福州 350108）

摘 要：中文句子评价对象抽取是指在中文句子中抽取评论所针对的对象或对象的属性。目前国内相关研究工作尚未能有效识别复合词评价对象和未登陆评价对象。针对以上两种情况，该文提出了一种基于层叠条件随机场的中文句子评价对象抽取方法。该方法首先通过低层条件随机场获得候选评价对象集，然后通过降噪模型对噪声进行过滤、补充模型对缺失的候选评价对象进行补充、合并模型对复合短语候选评价对象进行合并，最后由高层模型抽取出评价对象。实验结果显示，与基于线性链条件随机场的识别方法相比，该方法准确率、召回率和 F1 值分别提升 1.62%、5.75%和 4.17%，能有效地识别复合词评价对象和未登录评价对象，从而提高中文句子评价对象的识别精度。

关键词：评价对象；层叠条件随机场；降噪模型；补充模型

中图分类号：TP391 **文献标识码**：A

Identify Sentiment-Objects from Chinese Sentences Based on Cascaded Conditional Random Fields

ZHENG Minjie[1], LEI Zhicheng[2], LIAO Xiangwen[2], CHEN Guolong[2]

(1. College of Physics and Information Engineering, Fuzhou University, Fuzhou, Fujian 350108, China;
2. College of Mathematics and Computer Science, Fuzhou University, Fuzhou, Fujian 350108, China)

Abstract: Sentiment-objects extraction aims to identify the targets of opinion described in sentiment sentences. However, previous researches fail to extract compound targets and unknown words. In this paper, the cascaded CRFs model is presented to deal with the problem. The method first acquires opinion target set using lower-lever CRFs model. then, middle-lever models is employed to get candidate set by filtering noise, complementing missing candidate targets, and merging compound noun phrases. Finally, opinion targets set is extract from the higher-lever model using middle-lever model candidate set as input. Experiments show that our method outperforms linear chain CRFs by 1.62% in precision, 5.75% in recall, and 4.17% in F1 measure. Meanwhile, the method is also effective to identify the compound targets and unknown targets.

Key words: sentiment-objects; cascaded conditional random fields; noise reduction model; complement model

1 引言

近年来，随着博客、论坛、微博等网络媒介的迅猛发展，文本主观信息的抽取[1-2]逐步成为自然语言处理和信息检索等领域中的一个热点问题，而评价对象的抽取作为信息抽取中重要的一个研究课题，在电子商务、信息安全等领域具有重要的实用价值，引起了广泛关注。评价对象是指评论所针对的对象或对象的属性，如“笔记本电脑很方便。”这个观点句

收稿日期：2012-01-16 **定稿日期**：2012-03-19

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)；福州大学科技发展基金资助项目(2010-XQ-22)

作者简介：郑敏洁(1980—)，女，博士研究生，主要研究方向为无线通信与网络技术；雷志城(1987—)，男，硕士研究生，主要研究方向为文本倾向性检索与挖掘；廖祥文(1980—)，男，讲师，主要研究方向为文本倾向性检索与挖掘。

的评价对象是“笔记本电脑”,“方便”则是修饰“笔记本电脑”这一评价对象的评价短语(opinion expression)。如何准确全面地识别出中文句子的评价对象是一个难点问题。

关于评价对象抽取,国内外已经开展了很多的研究工作。其中 Li 等[3]、Xu 等[4]和 Zhu[5]等均通过构造启发式关联规则进行抽取。该方法由领域专家构造抽取规则,以进行模式匹配。构造的规则易于理解,但是很难保证规则的完备性和系统性,而且规则的领域相关性较高,系统的移植性较差。

此外,另外一种很重要的方法是基于自然语言处理(natural language processing, NLP)的方法。Hu 等[6-7]、刘鸿宇等[8]、Lu 等[9]、Ma 等[10]通过对语料进行语法分析进行抽取,Kim 等[11]在抽取句子成分时采用了语义角色标注的方法。NLP 的方法对于大规模的结构化文本测试效果较好,但在缺少语法结构或者语法结构复杂的非结构化文本和半结构化文本中表现则略显不足。

与以上两种方法相对的是基于统计模型的方法,该方法通过对抽取问题建立相应的数学模型进行抽取,根据所建立模型的自动化程度可分为非监督和监督两种。其中 Jin 等[12]通过自举方法实现语料的半自动标注,并使用 Lexical-HMM 分类器进行“观点实体”和“产品特征实体”的抽取;而宋晓雷等[13]则在模糊匹配并剪枝之后通过自举、聚类等方法等进行抽取;Qiu 等[14]采用双向传播(Double Propagation)的方法进行观点词的扩充和评价对象的抽取。非监督的方法无需人工标注大量语料,但是准确率有待提高。与非监督的机器学习方法相对的,监督的机器学习方法虽然需要事先对语料进行标注,但准确率较高,泛化能力较好。其中 Kim 等[15]、Somprasertsri 等[16-17]、章剑峰等[18]在抽取中采用了最大熵模型,Xia 等[19]则提出一种意见目标网络的方法用于提取名词术语。除此之外,评价对象抽取中经常采用条件随机场(Conditional Random Fields, CRFs)模型[20],该模型解决了最大熵等模型普遍存在的标记偏置问题(label bias problem),而且作为条件模型相对 HMM 等生成模型无需非常严格的独立性假设,可以灵活地引入多种特征。Jakob 等[21]在英文评价对象抽取中采用线性链结构的 CRFs 模型;针对中文的语料,徐冰等[22-23]、王中卿等[24]、张莉等[25]、Ding 等[26]均采用了线性链结构的条件随机场模型,并融合了词、词性、句法结构、本体知识等特征,取得了较好的结果。

综上所述,目前国内外研究中基于线性链 CRFs 模型的评价对象抽取方法取得较好的效果,但对于中文评价对象的抽取仍存在以下问题:

1) 当中文句子的评价对象是复合短语时,无法有效识别。复合短语是指评价对象经常嵌套多个名词、代词或动名词。评价对象是复合短语时,抽取较为困难。一方面,中文的词与词之间没有明显的边界标记符,这些被嵌套的词可能与上下文的词组合成复合词,造成抽取出的评价对象的边界不准确;另一方面,词性是判断评价对象的一个重要的特征,而中文的语法特点导致某些词的词性被误判,从而干扰评价对象抽取的效果。对于复合词评价对象的情况,线性链条件随机场无法准确地进行判断,经常只抽取出正确评价对象的一部分。

2) 评价对象中的未登录词情况无法很好处理。由于中文的特点及语料规模的限制,某些评价对象在语料中较少出现,导致在抽取过程中该词串判定为评价对象的权重低,无法有效抽取,导致部分句子无法识别出任何的评价对象。

针对以上问题,本文提出基于层叠条件随机场的中文句子评价对象抽取方法,以有效抽取中文复合名词评价对象及未登录评价对象。首先通过低层线性链条件随机场模型得到候选的评价对象集;然后针对候选评价对象集中复合词识别错误等问题通过降噪模型过滤处理,利用补充模型对因词语未登录等原因缺失候选任务评价对象的句子标识出一些可靠的候选评价对象,并通过合并模型对复合词候选评价对象进行合并;最后将处理之后得到的候选评价对象集输入到高层条件随机场模型,由高层模型识别出最终的评价对象。

本文主要结构如下:第 2 节介绍基于层叠条件随机场的评价对象抽取方法,第 3 节为实验结果与分析,第 4 节是结论。

2 基于层叠条件随机场的评价对象抽取方法

2.1 层叠条件随机场模型

条件随机场模型 CRFs 是 John Lafferty 和 Andrew McCallum[20]提出的一种无向图的模型,在中文分词、命名实体识别(Named Entity Recognition)、歧义消解等汉语自然语言处理任务中都有应用,并有着良好表现。但是对于复合词评价对象识

别精度差，对于未登录词识别效果存在缺陷。层叠条件随机场模型(Cascaded CRFs，CCRFs)按层叠加建立起多个层次的条件随机场模型，多个模型之间呈线性组合。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，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。其中刘康等[27]将层叠条件随机模型用于句子褒贬性的分析，而周俊生等[28]、杨晓东等[29]、郭剑毅等[30]在命名实体识别任务中也采用了该模型。层叠条件随机场模型如图 1 所示。

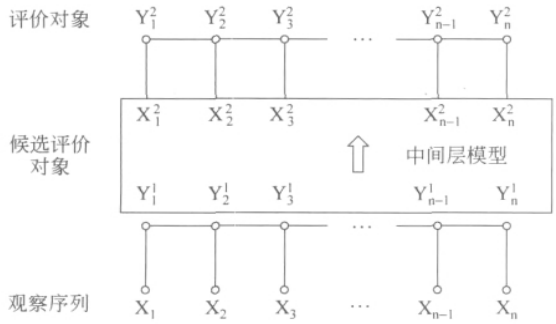

图 1　层叠条件随机场的无向图结构

低层模型中，$x=(x_1, x_2, \cdots, x_n)$为输入的观察序列，给定观察序列的情况下，输出序列 $y=(y_1^1, y_2^1, \cdots, y_n^1)$表示的是候选评价对象的序列，输出序列的条件概率为：

$$P(y \mid x)=\frac{1}{Z(x)}\exp\left(\sum_i\sum_k\lambda_k t_k(y_{i-1}, y_i, x, i)+\sum_i\sum_k\mu_k s_k(y_i, x, i)\right)$$

其中 $t_k(y_{i-1}, y_i, x, i)$，$s_k(y_i, x, i)$是特征函数，λ_k，μ_k是其对应的权重，由训练样本学习得到，$Z(x)$是归一化因子，定义如下：

$$Z(x)=\sum_y\exp\left(\sum_i\sum_k\lambda_k t_k(y_{i-1}, y_i, x, i)+\sum_i\sum_k\mu_k s_k(y_i, x, i)\right)$$

低层模型得到的候选评价对象 $y=(y_1^1, y_2^1, \cdots, y_n^1)$中的错误经过中间层模型的适当过滤和调整，得到 $y'=(y_1^{11}, y_2^{11}, \cdots, y_n^{11})$，结合其他特征高层 CRFs 模型的输入序列 $x^2=(x_1^2, x_2^2, \cdots, x_n^2)$，经过高层 CRFs 算法的处理输出得到评价对象的标记序列 $y^2=(y_1^2, y_2^2, \cdots, y_n^2)$，其条件概率为：

$$P(y^2 \mid x^2)=\frac{1}{Z(x^2)}\exp\left(\sum_i\sum_{k^2}\lambda_{k^2} t_{k^2}(y_{i-1}^2, y_i^2, x^2, i)+\sum_i\sum_{k^2}\mu_{k^2} s_{k^2}(y_i^2, x^2, i)\right)$$

$Z(x^2)$是其对应的归一化因子。

从高层标记序列 $y^2=(y_1^2, y_2^2, \cdots, y_n^2)$得到最终的评价对象的抽取结果。

2.2　特征选择

条件随机场模型一个重要的特点就是可以灵活地定义各种特征，用特征集合及其权重拟合样本的规律，以构建相应的模型。评价对象抽取的构成方式非常复杂，从评价对象的词性来看，大多数的评价对象是名词、代词或者名词短语，但也存在动词、从句等各种情况，因此本文考虑了词性特征。另一方面，在一些句子中，评价对象与相应的评价短语经常成对出现，而且通常存在着语法的依赖关系，如“诺基亚 N96 很炫”，评价对象“诺基亚 N96”与评价短语“炫”存在着语法上的依赖关系，故本文在引入上下文名词特征的同时考虑了语法依赖特征。针对上述特点，本文定义了如下的特征，如表 1 所示。

表 1　评价对象抽取采用的特征

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)，动词(v)
语法依赖	Dln	表征的是与评价短语存在直接依赖关系的词
上下文名词	Wrd	表征的是评价短语最为邻近的名词、代词或名词短语

词性特征有助于识别出名词、代词，为评价对象抽取提供词串之外的更多帮助信息。引入语法依赖特征可以识别与评价短语存在依赖的短语。实验过程中，由于语料中存在一个句子包含多个评价短语的情况，有些评价短语较长，包含多个 Token，造成与评价短语存在直接语法依赖的词串较多，产生噪声，所以本文在使用该特征时对一些情况进行了过滤，如连词、助词等较不可能是评价对象中的词。对于语法依赖特征本文进行如此表示：Ex 表示该 Token 是评价短语，Dln 表示该 Token 与评价短语存在直接语法依赖，no_Dln 表示不存在直接的语法

依赖。语法依赖可以有效引入某些评价短语与评价对象之间存在的关系，但由于很多情况评价短语与评价对象无直接依赖关系，因此利用评价对象有较大可能是名词、代词或复合名词这一特点，将离评价短语最近的名词、代词或名词短语标识出来作为一个特征 Wrd，本文使用 O_E 表示评价短语，nn_Noun 表示的是评价短语上下文中最为接近的名词（包含复合名词），other 表示其他。

对于层叠条件随机场，本文在低层模型和高层模型中采用同样的特征窗口，窗口大小均是[－3，3]，根据之前已开展的针对词、词性、句子倾向性、语法依赖关系、邻近名词等特征在中文评价对象抽取效用的研究，采用 Token＋Pos＋Dln＋Wrd 的特征组合抽取出的结果在准确率和召回率上都是最优的。因此，本文在低层模型中使用 Token＋Pos＋Dln＋Wrd 特征组合以获取候选评价对象，在高层模型中除了以上 4 个特征，本文将经过中间层处理的候选评价对象的识别结果作为一个特征输入。表 2 是模型采用的特征模板，其中 T_n 代表词串本身特征，P_n 代表词串的词性特征，D_n 表示词串的 Dln 特征，W_n 表示词串的上下文特征，L_n 表示词的标签。

表 2 评价对象抽取特征模板

序号	特征的符号化表示
1	T_nL_n；P_nL_n；D_nL_n；W_nL_n，$n\in[-3,3]$
2	$T_nL_{n-1}L_n$；$P_nL_{n-1}L_n$；$D_nL_{n-1}L_n$；$W_nL_{n-1}L_n$，$n\in[-3,3]$
3	$T_nT_{n+1}L_n$；$P_nP_{n+1}L_n$；$D_nD_{n+1}L_n$；$W_nW_{n+1}L_n$，$n\in[-3,2]$
4	$T_nT_{n+1}L_{n-1}L_n$；$P_nP_{n+1}L_{n-1}L_n$；$D_nD_{n+1}L_{n-1}L_n$；$W_nW_{n+1}L_{n-1}L_n$，$n\in[-3,2]$
5	$T_nT_{n+1}T_{n+2}L_n$；$P_nP_{n+1}P_{n+2}L_n$；$D_nD_{n+1}D_{n+2}L_n$；$W_nW_{n+1}W_{n+2}L_n$，$n\in[-3,1]$
6	$T_nT_{n+1}T_{n+2}L_{n-1}L_n$；$P_nP_{n+1}P_{n+2}L_{n-1}L_n$；$D_nD_{n+1}D_{n+2}L_{n-1}L_n$；$W_nW_{n+1}W_{n+2}L_{n-1}L_n$，$n\in[-3,1]$
7	$T_nT_{n+1}T_{n+2}T_{n+3}L_n$；$P_nP_{n+1}P_{n+2}P_{n+3}L_n$；$D_nD_{n+1}D_{n+2}D_{n+3}L_n$；$W_nW_{n+1}W_{n+2}W_{n+3}L_n$，$n\in[-3,0]$
8	$T_nT_{n+1}T_{n+2}T_{n+3}L_{n-1}L_n$；$P_nP_{n+1}P_{n+2}P_{n+3}L_{n-1}L_n$；$D_nD_{n+1}D_{n+2}D_{n+3}L_{n-1}L_n$；$W_nW_{n+1}W_{n+2}W_{n+3}L_{n-1}L_n$，$n\in[-3,0]$

2.3 中间层模型

本文在低层和高层模型中使用的是相同的训练语料，选择不同的特征，由此可能产生过拟合问题，因此我们在底层与高层模型中增加了中间层模型。由低层条件随机场识别出来的候选评价对象集，在复合词识别等方面存在错误，如果不进行处理，这些错误将输入到高层条件随机场模型中，经过高层模型扩散和传播，形成噪声，影响抽取的准确率。同时，低层条件随机场并未能完全抽取出全部的候选评价对象，而高层模型又依赖候选评价对象的抽取结果，这将导致高层模型无法抽取出更多正确的评价对象，影响抽取的召回率。基于上述考虑，本文提出以下 3 个中间层的处理模型：

2.3.1 降噪模型

该模型主要针对低层 CRFs 输出的候选评价对象集中复合词识别错误等进行过滤和调整，防止错误的扩散和传播。该模型主要基于以下规则：

1）过长候选评价对象的过滤：通过对语料的观察发现，评价对象的长度大部分在 1～3 个 Token 之间（分完词的结果），因此对于低层模型识别出的候选评价对象中长度超过一定阈值（4 个（不含 4 个）的 Token）的，统计其在语料中出现的次数，如其出现频率过低，则删除；否则保留。

2）过远候选评价对象的过滤：对于中文的句子，评价短语经常与评价对象存在一定关联，如评价对象一般与评价短语出现在同一个分句中，或者在相邻分句相近位置，虽然有时评价对象和评价短语距离较远，但这种情况较少。因此计算候选评价对象离评价短语的距离（有多个评价短语情况下取最近的一个），如果距离大于句子长度的一半，且该候选评价对象出现的次数小于一定阈值，则判定其为错误的候选评价对象，进行过滤。

3）标点的过滤：评价对象包含的标点主要有书名号，引号等，如《哈利波特》等电影名字等，对于除书名号、引号等之外的标点符号，如该标点符号单独出现（非小数点等情况），则进行过滤。

4）停用词的过滤：在中文句子中，有一些词成为评价对象的可能性极小，如“而且”、“了”（单独作为词串出现）等。本文构建了一个停用词表，将包含

这些停用词的候选评价对象作为噪声过滤。

5）评价短语的过滤：在标注时，评价短语与评价对象是独立存在的，因此评价短语作为候选评价对象的情况也作为噪声过滤。

2.3.2　补充模型

该模型主要针对某些句子中未识别出任何候选评价对象的情况，按照一定的规则自动标识出可能的候选评价对象。由于低层模型无法识别出所有句子的候选评价对象，同时经过降噪模型的处理，某些候选评价对象又被作为噪声排除，导致许多句子未能标识出任何候选评价对象，而候选评价对象的识别结果作为高层模型的一部分输入，又很大程度上影响了最后的识别结果，因此需要对无候选评价对象的句子进行处理。实验中对于分词后句子所包含的词数(token)进行统计得出句子的长度，对于长度不同的句子采取相应的处理规则，规则如下：

1）对于长度较长的句子(词数＞50)，句中的名词短语较多，如果全部作为候选评价对象，势必引入不必要的噪声，因此将句子中重复出现的名词短语标识为候选评价对象；如果不存在重复出现的名词短语，则将频率最高的名词短语标识为候选评价对象；

2）对于长度偏短的句子(词数＜10)，句中可能的候选评价对象较少，甚至没有，但它作为正确的候选评价对象的可能性也较高，因此将句子中所有的名词、短语均标识为候选评价对象，如果其中未含任何名词、代词和短语，则将与评价短语存在依赖关系且非评价短语的词串标记为候选评价对象；

3）对于其他句子(10＜词数＜50)，将与评价短语存在依赖关系而且离评价短语最近的名词、代词或名词性短语标识为候选评价对象，如果不存在同时满足这两种关系的名词、代词或名词性短语，则标识出评价短语最近的名词、代词或名词性短语，如果仍不存在，则标识出存在依赖的，否则置空。

2.3.3　合并模型

评价对象经常出现嵌套的现象，一个评价对象可能嵌套多个名词、代词或短语，即评价对象经常以复合词的形式出现。由于中文词与词之间无明显边界，加上分词工具本身存在误差，分词之后评价名词很可能与上下文的其他名词或短语形成复合词，出现分界错误，影响识别的准确率，而且复合词合并存在的错误很可能导致该复合词无法被识别为评价对象。因此，将识别出的候选评价对象进行 Token 之间的合并，即对于复合词的候选评价对象，将其词串组合成复合词，形成新的词串，对于新的词串，其词性为名词(n)，语法依赖特征及上下文名词特征则与合并前的一致。

3　实验结果与分析

3.1　语料

本文实验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带有倾向性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 3 所示。

表 3　语料分领域统计表

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　实验框架

实验中先对语料集进行分句、分词、词性分析、语法依存分析等预处理工作，经过低层条件随机场模型得到候选评价对象集，候选评价对象经过中间层处理之后的结果作为高层条件随机场模型的一部分输入，输入到高层条件随机场模型，得到最终的评价对象。本文分词和词性标注使用是中国科学院计算技术研究所提供的 ICTCLAS，评价短语直接使用答案中存在的评价短语，而候选评价对象的答案则由人工标注完成，语法依赖关系的分析使用的是 Stanford parser① 分析工具。

3.3　实验结果

本文中使用的是 CRF＋＋ 0.53 工具，其中的模型参数值，如-c -f -a 等，根据人工经验设定。实验中，为了减少人为因素的影响，采取是三倍交叉验证的方式，共进行 5 组对比实验：线性 CRFs，CCRFs(未经过中间层处理直接输入到高层模型)，CCRFs_降噪，CCRFs_降噪_补充，CCRFs_降噪_补充_合并。对于评价对象抽取的结果，本文采取严格的评价标准，只有抽取出的评价对象与答案完全匹

① http://nlp.stanford.edu/software/tagger.shtml

配才认为其是正确的，如评价对象答案是“笔记本电脑”，则“电脑”或“联想笔记本电脑”均被认为是错误的评价对象。本文的实验结果如表4所示。

表4 评价对象抽取评测结果对比/%

方法 \ 指标	P_准确率	R_召回率	F1值
CCRFs_降噪_补充_合并	60.32	50.85	55.18
CCRFs_降噪_补充	60.32	50.83	55.17
CCRFs_降噪	60.94	47.05	53.10
CCRFs	54.28	49.25	51.64
线性CRFs	58.70	45.10	51.01

1）对比实验结果可以看出CCRFs相对于线性CRFs在召回率方面大概提升了4%左右，但由于中间层没有规则过滤合并，造成低层CRFs模型的识别错误经过高层CRFs模型进一步放大，影响了抽取的准确性，使得准确率相对线性CRF降低了4.5%左右；

2）CCRFs_降噪在CCRFs的候选评价对象识别的基础上，对识别出的候选评价对象进行了基于一定规则的降噪过滤，相对于没有进行降噪处理的CCRFs虽然召回率有2.2%的下降，但准确率提高了6.66%左右，取得了60.94%的准确率，相对于线性CRF准确率提升了2.24%，召回率提升了1.95%，表5中为一些降噪模型过滤后的实例：

表5 降噪模型实例

方法 / 正确的评价对象	CCRFs	CCRFs_降噪
1. 单镜头/B 套机/I 2. 价格/B	而/O 宾得/B 的/I 单镜头/I 套/I 机/I 的/O 价格/B 不/O 高/O，仅/O 4000/O 元/O 出头/O 的/O 价格/B 是 家庭用户/O 的/O 绝佳/O 选择/O 。/O	而/O 宾得/O 的/O 单镜头/O 套/O 机/O 的/O 价格/B 不/O 高/O，仅/O 4000/O 元/O 出头/O 的/O 价格/B 是 家庭用户/O 的/O 绝佳/O 选择/O 。/O
1. 5400/B 转/I 硬盘/I 性能/I 表现/I	2710/B 配备/O 了/O 东芝/O 的/O 6008/O 硬盘/O，/O 具体/O 参数/O 为/O 18/O 英寸/O，/O 4200/O 转/O，/O 60/O，/O 该/O 磁盘/O 与/O 主流/O 的/O 5400/O 转/O 硬盘/O 性能/O 表现/O 存在/O 一定/O 的/O 差距/O，/O 可/O 升级/O 性/O 不及/O 主流/O 的/O 25/O 英寸/O 硬盘/O ./O	2710/O 配备/O 了/O 东芝/O 的/O 6008/O 硬盘/O，/O 具体/O 参数/O 为/O 18/O 英寸/O，/O 4200/O 转/O，/O 60/O，/O 该/O 磁盘/O 与/O 主流/O 的/O 5400/O 转/O 硬盘/O 性能/O 表现/O 存在/O 一定/O 的/O 差距/O，/O 可/O 升级/O 性/O 不及/O 主流/O 的/O 25/O 英寸/O 硬盘/O ./O

在第一个句子，“宾得 的 单镜头 套 机”是抽取出来的复合短语候选评价对象，因含5个token，被当作噪声过滤，虽然CCRFs_降噪未识别出“单镜头套机”这一评价对象，但避免了抽取出“宾得的单镜头套机”这一错误的评价对象。在第二个例子中，由于“2710”离评价短语“存在一定的差距”太远，依照规则进行了过滤，有效防止错误传入高层模型中。降噪模型可以有效地处理复合词评价对象识别存在的错误及其他处理对于提升准确率有着积极的作用；

3）CCRFs_降噪_补充相对于没有经过补充模型的CCRFs_降噪 提高了识别的召回率约3.8%，而准确率仅下降0.6%左右，虽然补充模型引入了一定的噪声，但其对召回率的提升作用是十分明显的。表6是经过补充模型处理的一些实例。

在第一个句子“外音喇叭保真度很差。”中没有识别出任何的候选评价对象，导致未能抽取出评价对象，按照补充规则将“外音喇叭保真度”标识出来后，高层模型顺利地抽取出这一评价对象。第二个句子中同样没有识别出任何的候选评价对象，而因为“体验”与评价短语“畅快”依赖，将“体验”标识为候选评价对象，最后由高层模型准确抽取出这一评价对象。补充模型有效地补足了因未登录词及其他原因未识别出的候选评价对象，对于更全面更好地抽取出未判别出来的候选评价对象是有积极意义的；

4）CCRFs_降噪_补充_合并相对CCRFs_降噪_补充多了合并候选评价对象的处理，但准确率并无提升，召回只上升0.2%，针对复合词评价对象的合并模型并未取得很好的结果，分析可能的原因如下：

表 6　补充模型实例

方法 / 正确的评价对象	CCRFs_降噪	CCRFs_降噪_补充
1. 外/B 音/I 喇叭/I 保真度/I	外/O 音/O 喇叭/O 保真度/O 很/O 差/O ./O	外/B 音/I 喇叭/I 保真度/I 很/O 差/O ./O
1. 体验/B	支持/O AV/O 输出/O 体验/O 更加/O 畅快/O：/O 影片/O、/O 歌曲/O、/O 图片/O 及/O 文本/O 可以/O 直接/O 输出/O 到/O 电视/O 上/O 播放/O ./O	支持/O AV/O 输出/O 体验/B 更加/O 畅快/O：/O 影片/O、/O 歌曲/O、/O 图片/O 及/O 文本/O 可以/O 直接/O 输出/O 到/O 电视/O 上/O 播放/O ./O

a. 合并候选评价对象时可以防止某些复合词的识别错误，但同时可能导致该复合词的候选评价对象的出现频率低，训练时权重过低，影响了抽取效果；

b. 合并候选评价对象之前虽然经过过滤，但候选评价对象本身仍存在一些分界的错误，而合成则造成错误的传播，造成一定的影响；而且不同的句子评价对象标注的不一致，如“诺基亚 N96 手机”在一些句子中评价对象是“诺基亚 N96”，而在另外的句子中可能是“诺基亚 N96 手机”，合并时“诺基亚 N96 手机”合并为一个复合词，造成一定程度上的影响；

c. 语料规模的影响，本文仅处理了 7 452 个观点句，训练不够充分，对抽取的结果造成影响；

综上所述，经过中间层降噪模型、补充模型和合并模型处理的 CCRFs_降噪_补充_合并取得了 F1 值 55.18%的结果，相对于线性链条件随机场模型提高了 4.17%。层叠条件随机场模型有效地处理了线性链条件随机场在复合词评价对象及未登录词等方面存在的问题，能够很好地应用于中文句子评价对象抽取任务。

4　总结

针对线性链条件随机场模型存在的不足，本文采用层叠条件随机场模型进行中文句子评价对象的抽取。通过采用降噪模型、补充模型和合并模型等中间层模型的过滤后，相对于线性链条件随机场准确率提升了 1.62%，召回率提升了 5.75%，F1 值提升了 4.17%，有效地抽取出了评价对象。

致谢

感谢中国科学院计算技术研究所为本文提供 ICTCLAS 分词工具。

参考文献

[1] James R Cowie, Wendy G Lehnert. Information extraction[J]. Communications of the ACM, 1996, 39(1): 80-91.

[2] Fuchun Peng, Andrew McCallum. Information extraction from research papers using conditional random fields[J]. Information Processing and Management, 2006, 42(4): 963-979.

[3] Li Zhuang, Feng Jing, Xiao-Yan Zhu. Movie review mining and summarization[C]//Proceedings of the ACM 15th Conference on Information and Knowledge Management. Arlington, Virginia, USA, 2006: 43-50.

[4] Ruifeng Xu, Chunyu Kit. Incorporating Feature-based and Similarity-based Opinion Mining-CTL in NTCIR-8 MOAT[C]//Proceedings of NTCIR-8 Workshop Meeting. Tokyo, Japan, 2010: 276-281.

[5] Shanzong Zhu, Yuanchao Liu, Ming Liu, et al. Research on Feature Extraction from Chinese Text for Opinion Mining[C]//Processing of 2009 International Conference on Asian Languages. Singapore, 2009: 7-10.

[6] Minqing Hu, Bing Liu. Mining Opinion Features in Customer Reviews[C]//Proceedings of 19th National Conference on Artificial Intelligence (AAAI-2004). California, USA, 2004: 755-760.

[7] Minqing Hu, Bing Liu. Mining and summarizing customer reviews[C]//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, Washington, USA, 2004: 168-177.

[8] 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.

[9] Bin Lu. Identifying Opinion Holders and Targets with Dependency Parser in Chinese News Texts[C]//Proceedings of the NAACL HLT 2010 Student Research

Workshop. Los Angeles, California, USA, 2010: 46-51.

[10] Tengfei Ma, Xiaojun Wan. Opinion Target Extraction in Chinese News Comments[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Poster Volume. Beijing, China, 2010: 782-790.

[11] Soo-Min Kim, Eduard Hovy. 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, Australia, 2006: 1-8.

[12] Wei Jin, Hung Hay Ho, Rohini K Srihari. OpinionMiner: 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, France, 2009: 1195-1204.

[13] 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.

[14] Guang Qiu, Bing Liu, Jiajun Bu, et al. Opinion Word Expansion and Target Extraction through Double Propagation[J]. Computational Linguistics, 2011, 37(1): 9-27.

[15] Soo-Min Kim, Eduard Hovy. Identifying Opinion Holders for Question Answering in Opinion Texts [C]//Proceedings of AAAI-05 Workshop on Question Answering in Restricted Domains. Pennsylvania, USA, 2005.

[16] Gamgarn Somprasertsri, Pattarachai Lalitrojwong. Automatic Product Feature Extraction from Online Product Reviews Using Maximum Entropy with Lexical and Syntactic Features[C]//Processing of The 2008 IEEE International Conference on Information Reuse and Integration. Las Vegas, Nevada, USA, 2008: 250-255.

[17] Gamgarn Somprasertsri, Pattarachai Lalitrojwong. A Maximum Entropy Model for Product Feature Extraction in Online Customer Reviews[C]//Processing of IEEE International Conference on Cybernetics and Intelligent Systems (CIS 2008). Chengdu, China, 2008: 575-580.

[18] 章剑锋,张奇,吴立德,等. 中文观点挖掘中的主观性关系抽取[J]. 中文信息学报,2008,22(2): 55-59.

[19] Yun-Qing Xia, Bo-Yi Hao, Liu-Ling Dai. Term Extraction from Web Reviews with Opinion Heuristics [C]//Proceedings of the Eighth International Conference on Machine Learning and Cybernetics. Baoding, China, 2009: 3516-3521.

[20] John D Lafferty, Andrew McCallum, Fernando C N Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [C]//Proceedings of the 18th International Conference on Machine Learning. Williamstown, MA, USA, 2001: 282-289.

[21] Niklas Jakob, Iryna Gurevych. 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, British Columbia, Canada, 2010: 1035-1045.

[22] 徐冰,王山雨. 句子级文本倾向性分析评测报告[C]//第二届中文倾向性分析评测会议(COAE2009) 论文集. 北京: 第二届中文倾向性分析评测委员会,2009: 69-73.

[23] 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): 1241-1247.

[24] 王中卿,王荣洋,庞磊等. Suda_SAM_OMS 情感倾向性分析技术报告[C]//第三届中文倾向性分析评测会议(COAE2011) 论文集. 北京: 第三届中文倾向性分析评测委员会,2011: 25-32.

[25] 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): 23-29.

[26] Shengchun Ding, Ting Jiang. Comment Target Extraction Based on Conditional Random Field & Domain Ontology[C]//Processing of 2010 International Conference on Asian Language. Harbin, Heilongjiang, China, 2010: 189-192.

[27] 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.

[28] 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.

[29] 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.

[30] 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.