文章编号:1009-3087(2013)02-0094-09

基于双层 HHMM 的产品评论特征和情感分类

张 磊,李梦诗,陈 黎,黎红友,李志蜀,彭 舰
(四川大学 计算机学院,四川 成都 610065)

摘 要:近年来,中文产品评论的特征情感分类是 Web 数据挖掘的重要研究内容之一。提出了一套完整的产品命名实体、特征词、情感词以及边界的标注规则,设计了多层次的混合标签模式;提出了双层 HHMM(层级隐马尔科夫模型)结构,将词形标注和词性标注的特点进行融合;提出了基于词形标注的 HHMM-1 算法和基于词性标注的 HHMM-2 算法,实现复杂短语的自动标注。实验证明,双层 HHMM 模型起到了互补的作用,模型的查全率和 F-score 值均有较大提高。

关键词:Web 数据挖掘;特征情感分类;标注规则;双层 HHMM
中图分类号:TP391 **文献标志码**:A

Features and Opinions Classification of Chinese Product Reviews Based on Two-level HHMMs

ZHANG Lei, LI Meng-shi, CHEN Li, LI Hong-you, LI Zhi-shu, PENG Jian
(College of Computer Sci., Sichua Univ., Chengdu 610065, China)

Abstract: In recent years, feature and opinion classification of Chinese product review is one of the most important research fields in Web data mining. A well-defined specification on data annotation for product named entities, features, opinions and boundaries was proposed and a hybrid tag representation was designed. By integrating linguistic features and POS features into automatic learning, a novel two-level Hierarchical HMMs (HHMMs) framework was put forward. The HHMM-1 and HHMM-2 algorithms were advanced to identify features and opinion entities automatically. The experimental results showed that two-level HHMM works in a mutual complementation way, which makes the recall and F-score of our approach obviously outstanding.

Key words: Web data mining; feature and sentiment classification; tagging specification; two-level HHMM

在当今飞速发展的 Internet 时代,“情感分析”或“观点挖掘”已经逐步成为 Web 数据挖掘领域的重点。Liu Bing 指出,产品特征和情感的提取与分类是情感分析最为关键的步骤[1-2]。

对于某一领域的产品评论进行特征级和短语级的情感分析[3],许多研究关注的是提取产品特征词(feature)及其相关联的情感词(opinion),形成特征情感对,从而可以判断基于特征级的情感[4]。

针对产品命名实体的识别(named entity recognition, NER),绝大多数研究都是基于监督学习[5]。Niu 等[6]使用隐马尔科夫模型和基于解析器的决策列表来识别产品特征,该方法难以准确地找到领域种子词汇,且系统性能依赖于解析器。Pierre[7]使用 Boolean 分类器,Bick[8]提出了基于规则的方法来实现 NER。但对于中文产品评论文档,还没有一个规范的和系统的标注规则。

针对特征和情感分类的研究,Hu、Turney 和 Pang[9-11]提出了基于统计的方法,利用关联规则识别高词频的特征词。Popescu 和 Etzioni[12]开发了无监督的信息提取系统 OPINE,使用松散标注方法识别倾向性。上述 2 类方法依赖于词的固定位置来发现词间的关联性,但针对语法结构灵活的中文表示则效果不佳。Zhuang、Jing 和 Zhu[13]通过提取高词频词来对电影评论进行分类和归纳,但由于使用的候选关键词列表固定不变,因而系统的识别能力有限。Ding、Liu 和 Yu[14-15]通过增加规则对 Hu 和 Liu[9,16]的方法进行扩展,但对边界的识别准确率较低,同时不能有效地识别低词频词。

收稿日期:2012-10-30
基金项目:四川省科技支撑计划资助项目(030405301054);四川大学青年教师科研启动基金资助项目(2011SCU11012)
作者简介:张 磊(1978—),男,讲师,博士. 研究方向:Web 数据挖掘;计算机网络;移动计算. E-mail: zhanglei@scu.edu.cn

根据上述研究的不足,以数码相机领域为研究背景,提出了一套完整的中文产品命名实体、特征词、情感词以及边界的标注规则;设计了双层 HHMM(hierarchical hidden markov model,层级隐马尔科夫模型);提出了基于词形标注的 HHMM-1 算法和基于词性标注的 HHMM-2 算法。

1　基于标注的双层 HHMM 算法

1.1　产品评论标注规则

1.1.1　边界和混合标签

根据 Liu 等[17]所持的观点,产品命名实体和特征情感词的识别可以设计为对已知词汇在实体中的位置进行正确标注的过程。一个已知词汇 w 会以 4 种形式表示[18],包括:

1)独立命名实体:〈〉;

2)命名实体的起始部分:〈BOE〉;

3)命名实体的中间部分:〈MOE〉;

4)命名实体的结束部分:〈EOE〉。

同时,一个混合的标签模式应该包含:命名实体、特征词、情感词和边界。根据 Liu 等[17]的原则,定义标签格式[18-19]如下:〈Tag_C - Tag_B〉,其中,Tag_C 代表了特征的类别标签,Tag_B 代表了特征的边界标签。

1.1.2　产品评论标注规则

根据 Zhao 和 Liu[20]的研究,在 HHMM 的训练中,需逐步找出特征情感的通用表达组合,以便更准确地识别产品命名实体和特征。同时,使用 Wei 等[21]提出的词汇组合扩展技术,用以自动收集和整理由专家所标注的特征词和情感词的相关词汇,然后进行扩展,构成特征词和情感词扩展集合。

涉及到的产品评论的特征标签总结如表 1 所示,详细的产品命名实体、特征和情感标签规范及示例请参考文献[18]。

1.2　双层 HHMM 模型架构

HMM 主要用于解决 3 类问题:评估、解码和学习[22]。主要针对"解码"问题进行研究。在对中文文本信息进行词语粗切分时,由于通用分词词典的局限性,一些领域内有意义的词汇往往被切分为没有意义的孤立词[23]。为了解决上述问题,通过研究发现,中文文本句子的表达结构都是以一定的语法规则来构成,典型的是:"动词 + 名词 + 形容词",如"拍摄效果好"。这样,可以对设定的标注规则进行训练,识别出语法结构的概率分布,即在产品评论文本所构成的观察序列中,自动标注出所需要的特征和情感,即状态序列。

表 1　产品评论的特征标签汇总

Tab. 1　Summary of product feature tags

产品评论特征标签	含义
〈Prod〉	产品名称
〈Prod_Bra〉	产品品牌
〈Prod_Typ〉	产品型号
〈Prod_Comp〉	产品部件
〈Prod_Func〉	产品功能
〈Prod_Feat〉	产品特征
〈Opinion_Pos_Exp〉	显式正面的情感
〈Opinion_Neg_Exp〉	显式负面的情感
〈Opinion_Pos_Imp〉	隐式正面的情感
〈Opinion_Neg_Imp〉	隐式负面的情感
〈BG〉	独立无关的词汇

由于 HHMM 比 HMM 功能更强,更适合于不同尺度、多层次的嵌套序列[24],根据 Fu 等[25]和岑咏华等[23]的研究,设计了一个基于双层 HHMM 的特征和情感分类系统架构,分为词形标注(命名实体、特征词和情感词以及边界标注)和词性标注(POS 标注)2 部分,分别对应上下 2 层相互关联的 HHMM 模型,即 HHMM-1 和 HHMM-2。如图 1 所示。

从上到下观察,共有 2 层 HHMM 模型按先后次序计算。上层 HHMM-1 中,输入文本粗分词结果,输出命名实体、产品特征和情感的最佳边界标记序列;下层 HHMM-2 中,利用上层模型输出作为输入,最终输出最佳词性标注序列。

从右到左观察,双层 HHMM 具有相同的结构,即在人工标注形成训练语料库之后,对测试语料进行自动标注。

系统主要由 5 个子模块组成:粗分词模块,用于命名实体和特征情感边界标注的模型参数训练模块,命名实体和特征情感边界标注模块,用于词性标注的模型参数训练模块及词性标注模块。

1.3　基于标注的双层 HHMM 算法

1.3.1　识别流程

基于 1.1 节所提出的标注原则,本双层 HHMM 算法的主要目的是,给定一个词汇序列及相应 POS 标注,从中找出最有可能的产品特征和情感的混合标签组合。

HHMM 模型分类过程如图 2 所示。

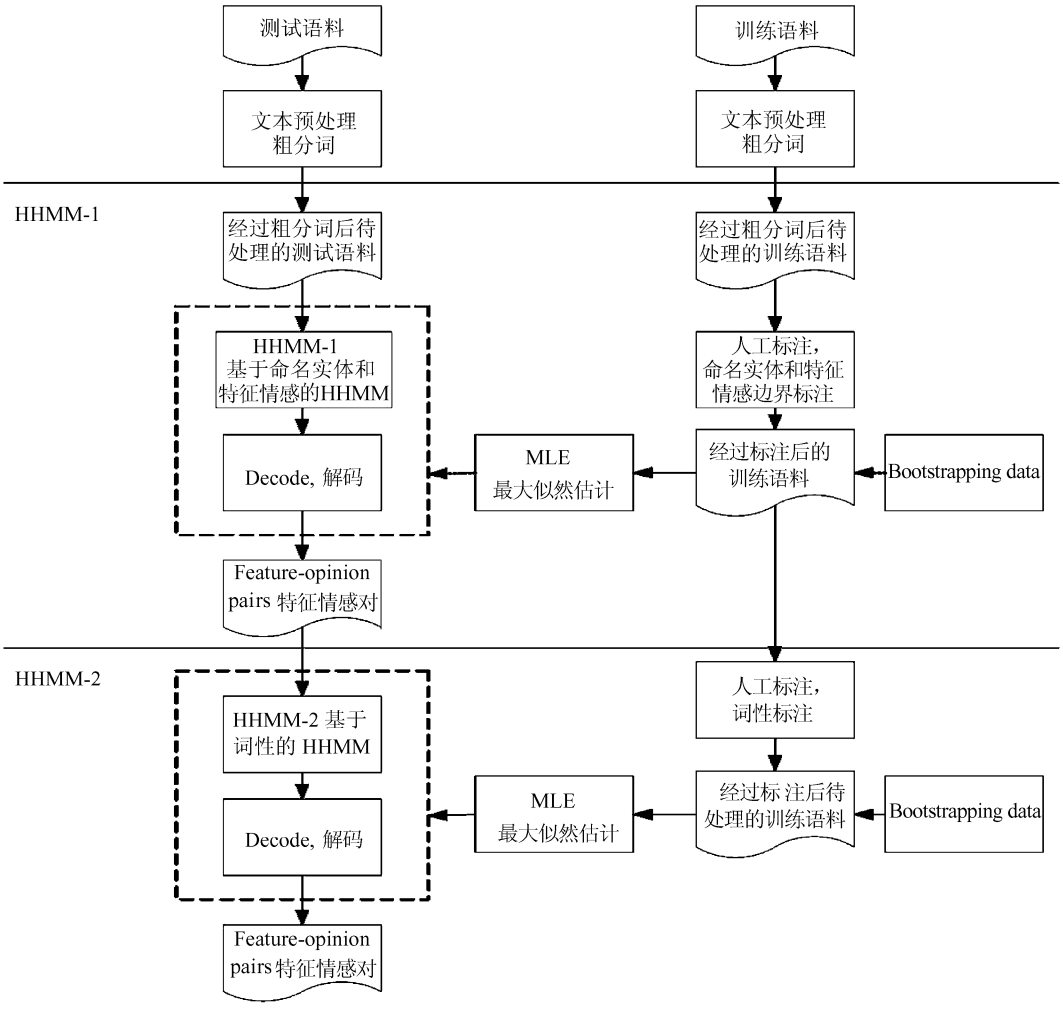

图 1　双层 HHMM 的产品评论特征和情感分类架构

Fig. 1　Two-level HHMMs framework for classifying the product-specific features and opinion entities

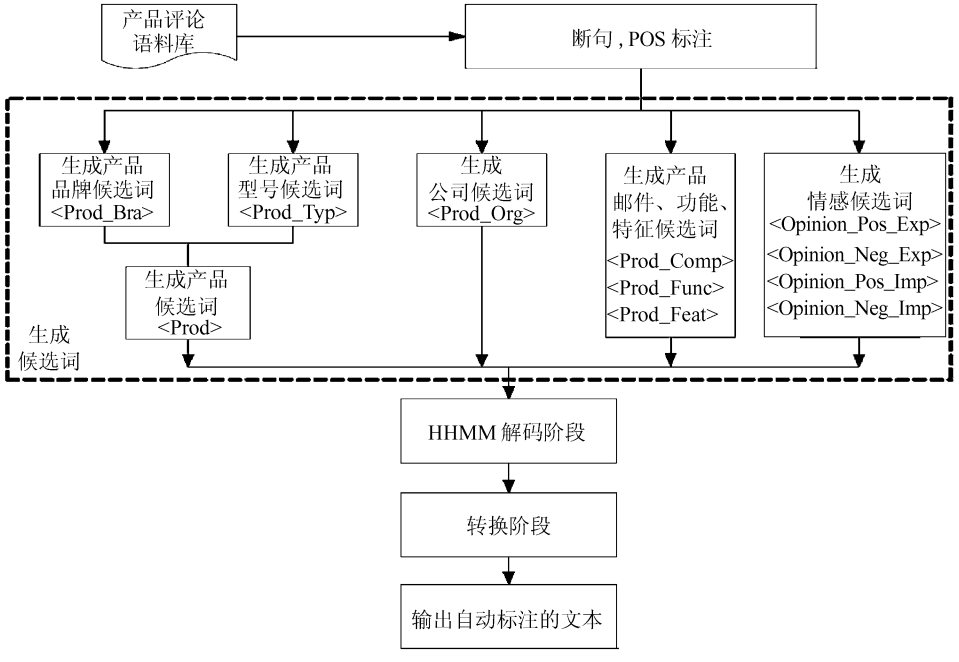

图 2　HHMM 模型识别的具体过程

Fig. 2　Tagging process of HHMM

该过程主要包含 3 个阶段:

1)候选标签的生成阶段:给定一个词汇序列及相应 POS 标注,生成候选的混合标签序列。这个阶段中,如何激活候选词的生成是最为关键的。如果激活条件过于松散,会产生大量的干扰;而如果条件过于严格,则又会大大降低查全率。因此,本模型使用了知识库方法和启发式方法的结合,详见文献[18]。

2)解码阶段:利用 Viterbi 算法遍历所有的候选混合标签序列,查找出满足最大似然估计的最优路径,即最优的标签序列。

3)转换阶段:为了便于更好地理解,所生成的混合标签序列将被简化。

接下来详细阐述双层 HHMM 的算法。

1.3.2　基于词形标注的 HHMM-1

1)基于标注的 HHMM-1 建模

将基本的 HHMM 算法进行扩展,设评论语句为一个词序列 $w_1/s_1, w_2/s_2, \cdots, w_n/s_n$,其中,$w_i$ 表示一个独立词,s_i 表示该词对应的POS标签,n 表示句子中词的个数。

模型中,POS 标签集包含了 2 部分内容:①根据北京大学词性标注集(PKU-POS)规定的基本词性标签集;②1.1 节提出的混合标签规则汇总。

构建如下的 HHMM 模型:

①状态集$\{S\}$,包含了:{ Prod, Prod_Bra, Prod_Typ, Prod_Comp, Prod_Func, Prod_Feat, Opinion_Pos_Exp, Opinion_Neg_Exp, Opinion_Pos_Imp, Opinion_Neg_Imp }和中文词汇表$\{V\}$。

②观察集$\{O\}$:同集合$\{V\}$相同。

在上述模型中,只有 Prod 作为转换过程中的内部状态,用以激活其它状态如 Prod_Bra, Prod_Typ 等,从而形成递归的 HHMM。基于 Fine 等[24]提出的理论,$q_i^d (1 \le d \le D)$ 用于表示在转换过程中,第 d 层的第 i 个状态。

因此,标注问题可以描述成:给定观察序列 $W = w_1 w_2 w_3 \cdots w_n$,基于动态拓扑结构变化的HHMM模型,找出最大概率的状态序列 Q^*,且该序列是一个多层次融合的状态序列。

根据贝叶斯理论,HHMM-1 定义如下:

$$\hat{T} = \arg\max_T P(Q \mid W) = \arg\max_T \frac{P(W \mid Q)P(Q)}{P(W)} = \arg\max_T P(W \mid Q)P(Q) \tag{1}$$

其中,$P(W) = 1$。从 HHMM 的根节点开始,中间状态的激活过程将按照转移概率,遍历不同层次的所有节点。为了描述这个过程,以第 k 层为例,其状态转换来自于第 $k-1$ 层的第 m 个状态。则转换概率 $P(Q)$ 和 $P(W \mid Q)$ 分别定义如下:

$$P(Q) \cong \underbrace{p(q_1^k \mid q_m^{k-1})}_{\text{垂直方向转换概率}} \times \underbrace{p(q_2^k \mid q_1^k) \prod_{j=3}^{|q^k|} p(q_j^k \mid q_{j-1}^k, q_{j-2}^k)}_{\text{水平方向转换概率}} \tag{2}$$

$$P(W \mid Q) = \begin{cases} \cong \prod_{j=1}^{|q_{PS}^k|} p([w_{q_j^k-begin} \cdots w_{q_j^k-end}] \mid q_j^k), \\ \quad \text{如果 } q_j^k \notin \{\text{中间状态}\}; \\ \text{激活下一层状态}, \\ \quad \text{如果 } q_j^k \in \{\text{中间状态}\} \end{cases} \tag{3}$$

式(2)和(3)中,q^k 代表第 k 层所有状态的数目,$|q_{PS}^k|$代表第 k 层所有与产品命名实体和特征情感相关的状态数目。

则在状态 q_j^k 时,词序表示如下:$w_{q_j^k-begin} \cdots w_{q_j^k-end}$。如果该词序表示产品特征情感及边界,则其标注如下例所示:

〈Prod_Comp-BOE〉对焦〈/Prod_Comp-BOE〉〈Prod_Comp-EOE〉系统〈/Prod_Comp -EOE〉。

为了描述命名实体和特征情感的识别过程,可以利用一个树形层次结构来表示 HHMM 的状态转移过程。示例句为:“佳能公司今年推出的佳能 EOS 50D,色彩还原十分不错”。其状态转移过程如图 3 所示[18]。

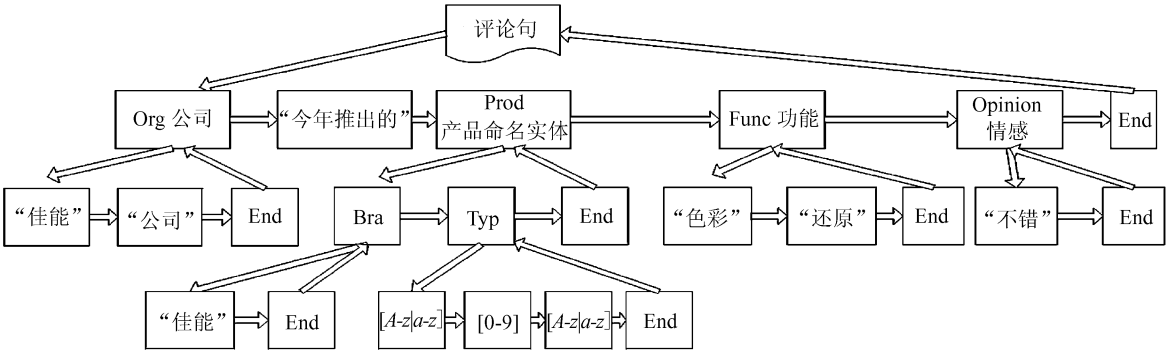

图 3　本文 HHMM 模型状态转移示例

Fig. 3　Sample of status switching in HHMM

2)命名实体和特征情感的自动标注

针对式(3),标注规则分析如下:

①若 $q_j^k \in \{V\}$:如"推出",则表示该词属于通用命名实体,说明预处理的结果正确。计算如下:

$$p([w_{q_j^k-begin}\cdots w_{q_j^k-end}] \mid q_j^k) = 1 \tag{4}$$

②若 $q_j^k \in \{\text{Prod}\}$:如"佳能 EOS 50D",由于 Prod 为内部状态,因此按照式(3),产品实体第 $k+1$ 层状态将会被激活,如此循环,直至到达结束标签位置。

③若 $q_j^k \in \{\text{Bra}\}$:如"佳能",由于该词既可能是 Bra 品牌候选词,也可能是 Org(公司或组织)候选词。则概率定义为:

$$p([w_{q_j^k-begin}\cdots w_{q_j^k-end}] \mid q_j^k = \text{Bra}) = p(q_1^{k+1} \mid q_j^k) = 0.5 \tag{5}$$

④若 $q_j^k \in \{\text{Typ}\}$:如"EOS 50D",以 PTT 规则[18]进行标注并加上复合标注。计算属于 Typ 的产品实体状态的发射概率,如式(6)所示。

$$p([w_{q_j^k-begin}\cdots w_{q_j^k-end}] \mid q_j^k = \text{Typ}) \cong p(ptt_1 \mid begin) \times p(end \mid ptt_{|q_j^k|}) \times \prod_{m=2}^{|q_j^k|} p(ptt_m \mid ptt_{m-1}) \tag{6}$$

其中,$|q_j^k|$表示当前状态下,第 k 层的观察序列的长度。

⑤若 $q_j^k \in \{\text{Feat},\text{Comp},\text{Func}\}$:如"色彩还原",参考 1.1.2 节提出的特征词扩展集,则以 Prod_Comp、Prod_Func 或者 Prod_Feat 进行标注并加上复合标注。

可以计算属于 Func 的产品实体状态的发射概率,如式(7)所示。

$$p([w_{q_j^k-begin}\cdots w_{q_j^k-end}] \mid q_j^k = \text{Func}) \cong p(func_1 \mid begin) \times p(end \mid func_{|q_j^k|}) \times \prod_{m=2}^{|q_j^k|} p(func_m \mid func_{m-1}) \tag{7}$$

其中,$|q_j^k|$意义同前。

同理,可以计算出属于 Comp 和 Feat 的产品实体状态的发射概率。

⑥若 $q_j^k \in \{\text{Opinion}\}$:如"不错",参考 Hownet 情感词典及情感扩展集,则当前词的标签以 OPINION_POS_EXP、OPINION_NEG_EXP、OPINION_POS_IMP 或 OPINION_NEG_IMP 进行标注。由于当前词汇属于情感词典,则定义如下:

$$p([w_{q_j^k-begin}\cdots w_{q_j^k-end}] \mid q_j^k) = 1 \tag{8}$$

3)命名实体和特征情感的自动提取

为了从上述模型中计算出最佳结果,在式(2)和(3)中,最为重要的 2 类概率是 $p(w_i \mid q_i)$ 和 $p(q_i \mid q_{i-1})$。$p(w_i \mid q_i)$ 表示状态序列 q_i 在观察序列 w_i 中的概率,$p(q_i \mid q_{i-1})$ 表示从状态 q_{i-1} 转移到 q_i 的转移概率。监督学习中,基于大规模语料库训练的前提,利用最大似然估计(maximum likelihood estimation, MLE)从训练语料对参数进行估计。具体计算如下:

$$p(w_i \mid q_i) = \frac{C(w_i,q_i)}{C(q_i)} \tag{9}$$

$$p(q_i \mid q_{i-1}) = \frac{C(q_{i-1},q_i)}{C(q_{i-1})} \tag{10}$$

其中,$C(w_i,q_i)$ 表示 w_i 作为状态 q_i 出现的次数,$C(q_i)$ 表示在整个状态序列中 q_i 出现的次数,$C(q_{i-1},q_i)$ 表示状态 q_{i-1} 下一个状态为 q_i 的次数。

这样,通过式(9)、(10),对状态序列进行训练和统计,即可得到命名实体和特征情感的状态字典和状态之间的转移概率。

1.3.3 基于词性标注的 HHMM-2

HHMM-2 和 HHMM-1 的区别主要在于状态集合 $S_{\text{Ⅱ}}$ 和观察集合 $O_{\text{Ⅱ}}$ 的不同。

设 HHMM-2 的观察序列 $T = t_1t_2t_3\cdots t_n$,则:

1)观察集合:$O_{\text{Ⅱ}} = \{\text{POS}\}$;

2)状态集合:$S_{\text{Ⅱ}} = \{\{\text{POS}\},$

$\{\text{Prod},\text{Prod\_Bra},\text{Prod\_Typ},\text{Prod\_Comp},\text{Prod\_Func},\text{Prod\_Feat},\text{Opinion\_Pos\_Exp},\text{Opinion\_Neg\_Exp},\text{Opinion\_Pos\_Imp},\text{Opinion\_Neg\_Imp}\}\}$。

则 HHMM-2 的模型定义如下:

$$\hat{Q}_{\text{Ⅱ}} = \arg\max_{Q_{\text{Ⅱ}}} P(Q \mid T) = \arg\max_{Q_{\text{Ⅱ}}} \frac{P(T \mid Q)P(Q)}{P(T)} = \arg\max_{Q_{\text{Ⅱ}}} P(T \mid Q)P(Q) \tag{11}$$

其中,$P(T) = 1$。HHMM-2 的具体计算类似于 HHMM-1,不再赘述。

1.3.4 上下层融合及解码

为了平衡区分性和鲁棒性,提高识别的整体性能,将 HHMM-1 和 HHMM-2 2 个模型进行融合,采用对数形式定义如下:

$$\overset{*}{Q} = \arg\max_{Q} P(Q \mid W,T) = \arg\max_{Q}\{\lg(P(Q)) + \lg(P(W \mid Q)) + \beta \times [\lg(P(Q)) + \lg(P(T \mid Q))]\} \tag{12}$$

其中:W 为 HHMM-1 的观察序列;T 为 HHMM-2 的观察序列;β 为可调参数,用以调节 2 个模型的权重值。

基于上述融合模型,则解码算法的主要目的是给定一个词汇序列,找出最大概率的同时满足词形和词性特征的产品评论混合标注序列。由于遍历所有路径的计算开销非常大,本模型使用 Viterbi 算法,即按照式(12),查找满足最大概率的最优路径。

Viterbi 算法的具体实现详见文献[18]。

2　实　验

2.1　数据准备

本实验评论数据来源于 ZOL 中关村在线(http://detail.zol.com.cn),利用作者已有研究[26],采用四川大学 Web 数据挖掘实验室的定向搜索引擎对数码相机板块进行爬取,重点针对专业性较强的产品评测文章和评论文章。

对于每个评论网页,从中提取出个人评论的文本内容,利用哈工大 HIT 的 LTP[27] 进行 POS 标注,形成评论文档语料库。

通过上述预处理过程,最终得到 1 727 篇评论文档。随机分为 2 个集合:1 个集合包含 1 435 篇文档,涉及到 10 种数码相机,作为训练集;另 1 个集合包含 292 篇文档,涉及到 6 种数码相机内容,作为测试集。

2.2　实验设计和评估方法

实验的评估标准采用经典的查准率(precision, p)、查全率(recall, r)和 F-score (F)。定义如下:

$$p = \frac{TP}{TP + FP}, r = \frac{TP}{TP + FN}, F = \frac{2pr}{p + r} \quad (14)$$

其中, TP 表示被正确分类的命名实体、特征词和情感词数量,FP 表示被错误分类的数量,FN 表示未被分类出的数量。

系统性能的评估主要针对系统自动标注的结果与人工标注的真实数据之间进行比较,只有 2 者完全匹配才认为是正确的识别结果。

为了进行对比,首先利用的是 Turney[10] 与 Hu 和 Liu[1] 的方法,实现一个 Baseline 系统,所用规则如表 2 所示。

表 2　Baseline 候选词提取规则

Tab. 2　Baseline for extracting candidate words

序号	第 1 个词	第 2 个词	第 3 个词
1	JJ	NN 或 NNS	任意
2	RB,RBR,RBS	JJ	NN 或 NNS
3	JJ	JJ	NN 或 NNS
4	NN 或者 NNS	JJ	不属于 NN, NNS

通过查找任意的名词和形容词来匹配表中的规则。如果满足规则,则提取为产品命名实体、特征或情感词,所在的句子则识别为观点句。

接下来,针对观点句识别以及情感倾向性判别,使用数码相机评论中情感倾向性比较明显的 20 个形容词作为情感词语或情感短语的核心种子词语,并利用 Wei 等[21] 的词汇组合扩展技术,查找所有种子词的同义词和反义词,从而构建词库。种子词集包括:

褒义种子集 Seed A = {快,先进,好,漂亮,美观,出色,稳定,方便,不错,满意},

贬义种子集 Seed B = {慢,低,差,遗憾,失望,落后,不足,粗糙,复杂,难看}。

2.3　实验结果和分析

2.3.1　双层 HHMM 融合参数 β 的估计

Pro、Typ 和 Bra 的识别率与 β 值的关系如图 4、5、6 所示,计算结果以百分比值进行对比。

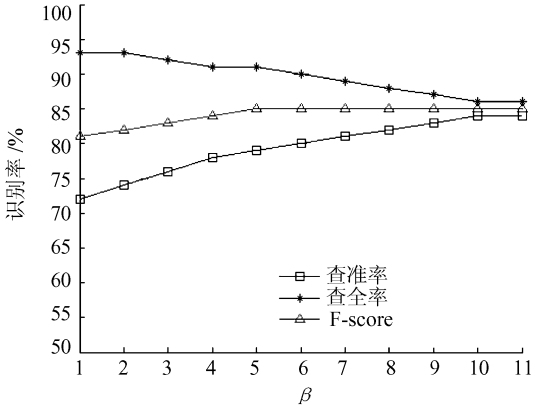

图 4　产品名称的识别率与 β 值的关系

Fig. 4　Recognition performance with the β value on product name

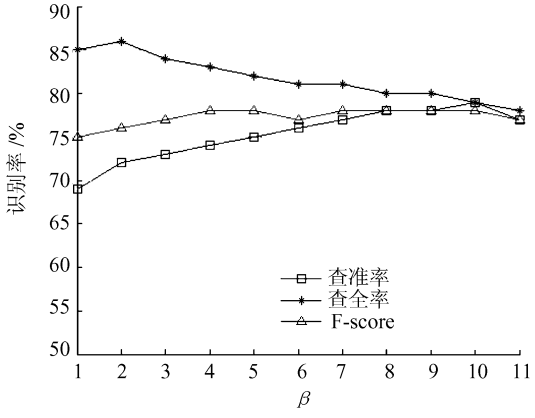

图 5　产品型号的识别率与 β 值的关系

Fig. 5　Recognition performance with the β value on product type

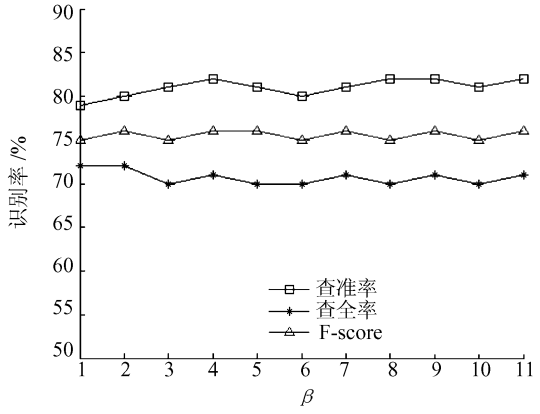

图6 产品品牌的识别率与β值的关系

Fig. 6 Recognition performance with the β value on product brand

Feat、Comp、Func 以及 Opinion 的识别率与β值的关系不再赘述。

从图4和5可以观察到，随着β值增大，Pro 和 Typ 的查准率和 F-score 从一开始逐步上升，并在一段时间后平衡，这表明了 HHMM-2 在一定程度上能够缓解数据稀疏问题带来的影响。但查全率一开始上升，之后逐步降低，反映了 HHMM-2 的不完备性。

与上述 Pro 和 Typ 变化明显不同的是 Bra 的变化，如图6所示。由于品牌的识别采用知识库的方法，且模型中参数的取值为常数，因此β值的影响很轻微，其查准率、查全率和 F-scores 的值变化很小。

从上述分析可以看出，HHMM-1 和 HHMM-2 起到了互补的作用，在本系统中，根据实验总结，β值一般设为6或者8。

2.3.2 识别性能对比

将本模型同 HHMM-1、HHMM-2、融合模型以及最大熵模型进行对比，其中，最大熵模型采用 Xiong 等[28]的方法，训练了2层最大熵模型，使用内层识别产品型号，外层识别产品名称。结果见表3和4。

表3 产品名称、型号识别性能比较(β=8)

Tab. 3 Performance comparison for extracting product name, type and brand

模型	产品名称			产品型号		
	查准率/%	查全率/%	F-score/%	查准率/%	查全率/%	F-score/%
HHMM-1	62.1	83.2	70.3	69.3	93.2	79.4
HHMM-2	82.3	69.7	74.1	92.1	77.6	84.2
HHMM-1 + HHMM-2	77.5	80.5	78.7	83.5	89.3	85.4
最大熵	81.2	59.4	68.2	82.8	43.2	56.3

表4 产品特征、功能识别性能比较(β=8)

Tab. 4 Performance comparison for extracting feature, component and function

模型	产品特征			产品功能		
	查准率/%	查全率/%	F-score/%	查准率/%	查全率/%	F-score/%
HHMM-1	83.4	71.1	76.1	51.2	74.1	60.5
HHMM-2	85.4	68.8	77.2	55.5	72.5	61.7
HHMM-1 + HHMM-2	87.3	72.5	79.3	64.2	87.2	74.1
最大熵	87.6	61.1	71.9	85.9	75.8	80.5

其它内容，如产品品牌和产品部件的识别性能对比，可参考文献[18]，这里不再赘述。

可以看出，3个 HHMM 模型的性能在查全率和 F-score 上均比最大熵模型优异。这是由于最大熵模型的处理过程总是按照输入序列从左到右进行，这种顺序的识别模式使得识别的错误不断叠加，不同层次之间的信息也无法进行互补。而作者提出的双层 HHMM，能够在每个层次增加判断条件，这对于识别嵌套结构以及长度可变的复杂结构的文本特别有效。

上述结果中，HHMM-1 的查准率较低，而查全率较高，反过来 HHMM-2 的查准率较高，而查全率较低，充分说明只有融合模型才能够 HHMM-1 和 HHMM-2 的优点结合起来，从而得到较高的 F-score 值。

同时，利用表2规则提取观点句，使用文献[18]的情感倾向性判定算法进行情感极性判断，结果见表5。可以看出，作者提出的 HHMM 相比于 Baseline，无论是查准率、查全率还是 F-score 值，都有很大的提高，特别在复杂短语的识别上比较突出。

表5 观点句提取和极性识别性能比较(β=8)

Tab. 5 Performance comparison for opinion recognition and polarity classification

模型	观点句提取			命名实体和特征的情感极性		
	查准率/%	查全率/%	F-score/%	查准率/%	查全率/%	F-score/%
HHMM-1	85.7	80.8	83.2	68.7	65.9	67.1
HHMM-2	88.7	79.5	83.9	70.9	65.3	67.3
HHMM-1 + HHMM-2	89.1	80.8	84.8	74.0	66.6	70.5
最大熵	85.5	69.1	76.5	73.3	60.0	66.1
Baseline	46.3	57.0	51.1	13.2	20.7	16.1

3　结论和展望

提出了双层 HHMM 的产品评论特征和情感分类模型,设计了多层次的混合标注规则,提出了 HHMM-1 和 HHMM-2 算法,并将 2 层模型进行融合解码,通过实验证明了该模型的性能相比传统模型有较大提高。

由于中文语法结构的复杂性和灵活性,提出的特征提取和情感分析模型在处理一些复杂语句的时候,查准率不是很高,在未来的研究中,应采用更广泛的语义标注方法加工更大规模的语料,进一步探索语义和语法信息在情感分析中的作用。

参考文献:

[1] Hu Minqing, Liu Bing. Mining opinion features in customer reviews[C]//Proceedings of AAAI. Washington D C: AAAI Press, 2004: 755 – 760.

[2] Liu Bing. Web data mining: Exploring hyperlinks, contents, and usage data[M]. New York: Springer, 2006.

[3] Yu Hong, Hatzivassiloglou V. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2003). 2003: 1777 – 1783.

[4] Nasukawa T, Yi J. Sentiment analysis: Capturing favorability using natural language processing[C]//Proceedings of the Conference on Knowledge Capture. 2003: 61 – 67.

[5] Yi J, Nasukawa T, Bunescu R, et al. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques[C]//Proceedings of the IEEE International Conference on Data Mining (ICDM 2003). 2003: 379 – 386.

[6] Niu C, Li W, Ding J, et al. A bootstrapping approach to named entity classification using successive learners[C]//Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. 2003: 335 – 342.

[7] Pierre J M. Mining knowledge from text collections using automatically generated metadata[C]//Proceedings of Fourth International Conference on Practical Aspects of Knowledge Management (PAKM2002). 2002: 537 – 548.

[8] Bick E. A named entity recognizer for Danish[C]//Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004). 2004: 305 – 308.

[9] Hu M, Liu B. Mining and summarizing customer reviews[C]//Proceedings of the 10th ACM International Conference on Knowledge Discovery and Data Mining (KDD'04). 2004: 168 – 177.

[10] Turney P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews[C]//Proceedings of the Association for Computational Linguistics (ACL 2002). 2002: 417 – 424.

[11] Pang Bo, Lee L. Opinion mining and sentiment analysis[J]. Foundations and Trends in Information Retrieval, 2008, 2(1): 1 – 135.

[12] Popescu A-M, Etzioni O. Extracting product features and opinions from reviews[C]//Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005). 2005: 339 – 346.

[13] Zhuang L, Jing F, Zhu X. Movie review mining and summarization[C]//Proceedings of the International Conference on Information and Knowledge Management. 2006: 43 – 50.

[14] Ding X, Liu B, Yu P S. A holistic lexiconbased approach to opinion mining[C]//Proceeding of the International Conference on Web Search and Web Data Mining (WSDM'08). 2008: 231 – 239.

[15] Liu Bing, Hu Minqing. Opinion observer: Analyzing and comparing opinions on the web[C]//Proceedings of WWW 2005. 2005: 1195 – 1203.

[16] Ding X, Liu B, Yu P S. A holistic lexiconbased approach to opinion mining[C]//Proceeding of the International Conference on Web Search and Web Data Mining (WSDM'08). 2008: 231 – 239.

[17] Liu Feifan, Zhao Jun, Lv Bibo, et al. Product named entity recognition based on hierarchical hidden Markov model[C]//Proceedings of the ACL Fourth SIGHAN Workshop. 2005.

[18] Zhang Lei. Research on the key technologies of WEB data mining in commerce[D]. Chengdu: Sichuan University,

2011.[张磊.商业 WEB 挖掘关键技术研究[D].成都:四川大学,2011.]

[19]Wang Ke,Liu Yuan,Luo Wanbo,et al. Analysis of network information based on Chinese text topic tracking[J]. Journal of Sichuan University:Engineering Science Edition,2004,36(1):114 - 118.[王科,刘渊,罗万伯,等.基于中文文本主题跟踪的网络信息分析[J].四川大学学报:工程科学版,2004,36(1):114 - 118.]

[20]Zhao Jun,Liu Feifan. Product named entity recognition in Chinese text[J]. Journal of Language Resource and Evaluation,2008,2(2):197 - 217.

[21]Wei Jin,Hung H H. A novel lexicalized HMM-based learning framework for web opinion mining[C]//Proceedings of the 26th International Conference on Machine Learning. 2009:465 - 472.

[22]Seymore K,McCallum A,Rosenfeld R. Learning hidden markov model structure for information extraction[C]. AAAI 99 Workshop on Machine Learning for Information Extraction,1999:37 - 42.

[23]Cen Yonghua,Han Zhe,Ji Peipei. Chinese term recognition based on hidden markov model[J]. New Technology of Library and Information Service,2008,24(12):54 - 58.[岑咏华,韩哲,季培培.基于隐马尔科夫模型的中文术语识别研究[J].现代图书情报技术,2008,24(12):54 - 58.]

[24]Fine S,Singer Y,Tishby N. The hierarchical hidden Markov model:Analysis and applications[J]. Machine Learning,1998,32(1):41 - 62.

[25]Fu G,Luke K. Chinese named entity recognition using lexicalized HMMs[J]. ACM SIGKDD Explorations Newsletter,2005,7(1):19 - 25.

[26]Zhang Lei,Li Zhishu,Wang Yongfeng,et al. Lead user opinion consensus based on Deffuant model[J]. Computer Integrated Manufacturing Systems,2011,17(10):2101 - 2111.[张磊,李志蜀,王永锋,等.基于 Deffuant 模型的领先用户意见一致性研究[J].计算机集成制造系统,2011,17(10):2101 - 2111.]

[27]Che Wanxiang,Li Zhenghua,Liu Ting. LTP:A Chinese language technology platform[C]//Proceedings of the Coling,2010. 2010:13 - 16.

[28]Xiong D,Yu H,Liu Q. Tagging complex NEs with Maxent models:Layered structures versus extended Tagset[C]//Proceedings of the First International Joint Conference on Natural Language Processing(IJCNLP-04). 2004:638 - 643.

(编辑 杨 蓓)